Statistical Inference Analysis on Student Performance

MASOUD RAYAT ZADEH

Jan 17th 2024

Instructions to students

Save this template as your studentID.Rmd; you will upload this file as part of your submission. Change the author information on line 3 of this file to your **student ID**. Do not change the authorship to your name.

Your should knit this file to a document **Word** format. The Word document is what will be marked!

Any changes that you make to the data (e.g. variable name changes) should be made entirely within R.

The subsubsections labelled **Answer:** indicate where you should put in your written Answers. The template also provides blank code chunks for you to complete your Answers; you may choose to add additional chunks if required.

This is an individual assessment: do not work with any other person during this exam. Text-matching software will be used on all submissions.

Instructions for submission

You must submit your assignment before the stated deadline by electronic submission through Blackboard.

- It is a good idea to save your work early and frequently to ensure you have no issues with the submission portal. Multiple submissions can be made to the portal, but only the final one will be accepted.
- It is your responsibility to submit the exam in a format stipulated above. Your marks may be affected if your tutor cannot open or properly view your submission.
- Do not leave submission to the very last minute. Always allow time in case of technical issues.
- The date and time of your submission is taken from the Blackboard server and is recorded when your submission is complete, not when you click Submit.

- It is essential that you check that you have submitted the correct file(s), and that each complete file was received. Submission receipts are accessed from the Coursework tab.

There is no late submission permitted on this timed assessment. Ensure that you submit your submission in good time. Neither the module leader nor module team can accept late assessments, do not ask them to do so.

Background to the research

The head of school for a four year degree course has provided you with some data based on student demographics, marks and graduate outcomes.

They have asked you if the data could reveal findings that may be relevant for monitoring student performance and outcomes.

Data instructions

Your individual data set is accessed via Blackboard >>> Assessments >>> Dewis Data For Exam.

You must only analyse the specified data. No other data is to be used for this assessment.

All data manipulation and analyses must be done within R.

Data structure

The variables collected for each student are:

studentID – a unique student identifier issued to each student at the start of the course

outcome – employment status one year after finishing the course (E1 = employed in a graduate role, E2 =Employed in a non-graduate role, Education = in full time further education, Unemployed = not yet employed)

age – age at start of course

gender – gender at start of course

language – score given for student level of English proficiency determined as part of the application process for the course (minimum 0, maximum 10)

feedback – score given by student for their satisfaction of the course when asked at the end of Year 4 (minimum 0, maximum 10)

Mark1 - Mark for Year 1 (out of 100)

Mark2 - Mark for Year 2 (out of 100)

Mark3 - Mark for Year 3 (out of 100)

Mark4 – Mark for Year 4 (out of 100)

QUESTIONS START HERE

Question 1: Data Preparation

a) Ensure you have prepared your knitted Word document as per Instructions to Students

b) You should load the data in R, describe and perform any actions with respect to:

-any manipulation of the data structure

-missing values

-ensuring data is valid

(10 marks)

Answer:

I loaded the data from the specified path into a dataframe named contents using the read.csv function. To get an initial overview of the dataset, I displayed the first few rows using the head function. This step ensured the correct loading of the data and enabled a preliminary examination of its structure and contents.

I began the data cleaning process by conducting an initial exploration of the entire dataset through summarization. After reviewing the summary and dataset structure, I decided on specific data cleansing approaches to prepare the data for analysis.

For handling missing values in the “language” and “feedback” columns, I chose to impute them with the mean value of each respective column. This approach was favoured over other methods, such as removing rows with missing values, primarily because a significant number of rows contained NA values, and I aimed to utilise the available data values on these rows. Additionally, both “language” and “feedback” are continuous variables, making mean imputation a suitable choice.

The “age” variable did not have a specified acceptable range, so I assumed that individuals of all ages could participate in this course and set an upper limit of 100. Upon filtering the data, I found only one observation that exceeded this range. In such cases, removing the outlier is a practical solution since the proportion of missing values is small and distributed randomly.

To ensure the dataset’s cleanliness and readiness for analysis, I employed the “assert” and “verify” functions. This systematic data cleaning approach aimed to enhance the dataset’s quality and reliability, setting a solid foundation for subsequent analyses.

```r
# load the dataset here
student_performance <- read.csv("D:/Students-Performance-
Analysis/data/contents.csv")

# further data preparation here

# Exploring the data
summary(student_performance)
```

```
   studentID        gender               age            outcome
 Min.   :11607   Length:233         Min.   :  18.00   Length:233
 1st Qu.:30265   Class :character   1st Qu.:  18.00   Class :character
 Median :55070   Mode  :character   Median :  18.00   Mode  :character
 Mean   :55363                      Mean   :  29.65
 3rd Qu.:79541                      3rd Qu.:  20.00
 Max.   :99908                      Max.   :2001.00

    language        feedback         Year1           Year2           Year3
 Min.   :2.000   Min.   :0.000   Min.   :49.00   Min.   :53.00   Min.   :67.0
 1st Qu.:5.000   1st Qu.:2.000   1st Qu.:52.00   1st Qu.:59.00   1st Qu.:73.0
 Median :6.000   Median :3.000   Median :54.00   Median :61.00   Median :74.0
 Mean   :5.714   Mean   :2.991   Mean   :53.73   Mean   :60.61   Mean   :74.2
 3rd Qu.:6.000   3rd Qu.:4.000   3rd Qu.:55.00   3rd Qu.:62.00   3rd Qu.:76.0
 Max.   :9.000   Max.   :6.000   Max.   :59.00   Max.   :66.00   Max.   :81.0
 NA's   :20      NA's   :15
     Year4
 Min.   :63.00
 1st Qu.:69.00
 Median :70.00
 Mean   :70.29
 3rd Qu.:72.00
 Max.   :75.00
```

```r
str(student_performance)
```

```
'data.frame':   233 obs. of  10 variables:
 $ studentID: int  34166 22525 87848 83664 64568 59203 32657 17882 23047
19867 ...
 $ gender   : chr  "Male" "Male" "Male" "Male" ...
 $ age      : int  18 18 18 19 19 18 18 18 18 22 ...
 $ outcome  : chr  "E1" "E1" "E2" "Unemployed" ...
 $ language : int  7 6 6 6 6 5 6 4 6 7 ...
 $ feedback : int  4 5 4 1 3 3 3 2 3 4 ...
 $ Year1    : int  56 59 54 50 54 53 57 55 54 56 ...
 $ Year2    : int  62 63 59 57 61 64 62 60 60 62 ...
 $ Year3    : int  75 78 75 68 76 74 75 75 74 74 ...
 $ Year4    : int  71 72 71 63 71 74 68 70 69 70 ...
```

```r
# Count missing values in the "language" and "feedback" columns
missing_values_lang <- sum(is.na(student_performance$language))
```

```r
missing_values_feed <- sum(is.na(student_performance$feedback))
cat("Number of missing values in 'language' column:", missing_values_lang, "
")
```

```
Number of missing values in 'language' column: 20
```

```r
cat("Number of missing values in 'feedback' column:", missing_values_feed, "
")
```

```
Number of missing values in 'feedback' column: 15
```

```r
# Checking non-numeric columns
table(student_performance$gender)
```

```
Female   Male
   103    130
```

```r
table(student_performance$outcome)
```

```
        E1         E2  Education Unemployed
        77         41         27         88
```

```r
# Check for duplicates in the studentID column
duplicate_count <- sum(duplicated(student_performance$studentID))
cat("Number of duplicates in 'studentID' column:", duplicate_count, "
")
```

```
Number of duplicates in 'studentID' column: 0
```

```r
# Replace NA values with mean of the column
student_performance$feedback[is.na(student_performance$feedback)] <-
round(mean(student_performance$feedback, na.rm = TRUE), 0)
student_performance$language[is.na(student_performance$language)] <-
round(mean(student_performance$language, na.rm = TRUE), 0)

# Remove out-of-range ages
student_dataset <- subset(student_performance, age < 100)
rm(student_performance)

# Final data check
student_dataset %>%
  verify(has_all_names("studentID", "gender", "age", "outcome", "language",
"feedback", "Year1", "Year2", "Year3", "Year4")) %>%
  assert(not_na, studentID, gender, age, outcome, language, feedback, Year1,
Year2, Year3, Year4) %>%
  verify(age >= 10) %>%
  verify(age <= 100) %>%
  assert(in_set("Male", "Female"), gender) %>%
  assert(in_set("E1", "E2", "Unemployed", "Education"), outcome) %>%
  assert(within_bounds(10000, 99999), studentID) %>%
```

```
  assert(within_bounds(0, 10), language) %>%
  assert(within_bounds(0, 10), feedback) %>%
  assert(within_bounds(0, 100), Year1) %>%
  assert(within_bounds(0, 100), Year2) %>%
  assert(within_bounds(0, 100), Year3) %>%
  assert(within_bounds(0, 100), Year4)
```

	studentID	gender	age	outcome	language	feedback	Year1	Year2	Year3	Year4
1	34166	Male	18	E1	7	4	56	62	75	71
2	22525	Male	18	E1	6	5	59	63	78	72
3	87848	Male	18	E2	6	4	54	59	75	71
4	83664	Male	19	Unemployed	6	1	50	57	68	63
5	64568	Male	19	Unemployed	6	3	54	61	76	71
6	59203	Male	18	E1	5	3	53	64	74	74
7	32657	Male	18	Education	6	3	57	62	75	68
8	17882	Male	18	Education	4	2	55	60	75	70
9	23047	Male	18	E1	6	3	54	60	74	69
10	19867	Male	22	E2	7	4	56	62	74	70
11	34834	Female	19	E1	6	3	54	64	74	72
12	80562	Male	18	Unemployed	5	3	54	61	73	69
13	37597	Female	18	E1	6	3	55	62	77	70
14	53648	Female	18	Unemployed	4	0	51	59	70	67
15	35361	Female	18	E2	5	2	52	62	71	71
16	88832	Female	20	Unemployed	4	1	51	60	71	68
17	73196	Male	19	Unemployed	7	2	50	58	73	67
18	94002	Male	18	E2	9	4	56	63	75	69
19	58435	Male	18	Education	6	2	49	61	72	70
20	39335	Female	18	Unemployed	5	1	53	57	74	65
21	35601	Male	18	E1	5	3	56	62	77	73
22	70085	Male	18	E2	5	3	51	58	72	70
23	66579	Male	19	Education	5	2	52	60	72	69
24	95145	Male	18	Unemployed	6	3	53	59	72	69
25	28089	Male	20	E1	6	4	56	62	77	72
27	11824	Female	18	E2	6	3	52	61	72	71
28	14559	Male	20	Unemployed	6	3	56	61	74	70
29	50581	Female	21	Unemployed	6	4	54	62	72	71
30	69601	Female	18	Unemployed	5	2	51	57	72	67
31	91217	Male	20	Unemployed	7	2	53	61	74	69
32	51420	Male	20	E2	7	4	53	61	74	71
33	19158	Male	19	Unemployed	6	4	54	63	74	72
34	48021	Male	20	Unemployed	2	1	54	59	73	71
35	66015	Male	18	E1	8	5	55	64	75	71
36	25673	Female	20	E1	5	3	58	65	77	73
37	39668	Male	18	E1	7	3	53	61	73	71
38	21189	Male	18	E1	4	1	53	61	74	69
39	98818	Male	22	Education	6	3	53	62	72	70
40	32594	Male	19	Unemployed	7	3	54	59	76	71
41	88821	Female	18	E1	5	5	55	62	76	75
42	91614	Female	19	E1	6	4	56	66	75	73
43	42558	Female	18	E1	6	3	53	59	73	71

44	40572	Male	18	E1	6	5	55	61	76	73
45	37700	Male	18	E2	6	4	55	60	74	69
46	89656	Male	21	E1	7	4	55	63	77	72
47	15984	Male	18	Unemployed	3	3	53	62	73	71
48	34125	Male	18	E1	6	4	56	63	75	73
49	30265	Female	19	E2	6	3	53	61	74	70
50	23085	Male	18	Unemployed	7	1	51	56	74	65
51	63607	Male	18	E2	3	2	56	61	72	69
52	74122	Male	18	E2	7	3	52	60	76	71
53	82387	Male	19	Unemployed	4	1	52	58	74	70
54	62074	Female	18	E1	4	3	55	62	74	74
55	59134	Female	21	Unemployed	6	2	53	62	73	69
56	32485	Female	18	Unemployed	6	3	52	59	74	71
57	84713	Male	18	Unemployed	6	2	52	58	74	67
58	20970	Male	18	Unemployed	5	2	52	58	74	71
59	71277	Female	19	Unemployed	6	3	52	58	73	68
60	95483	Female	18	E1	8	5	54	63	75	70
61	43010	Male	18	Unemployed	6	3	55	62	74	70
62	40978	Female	18	Unemployed	4	2	53	59	71	69
63	32488	Male	18	E2	6	2	53	61	76	71
64	39654	Female	19	E1	5	2	54	59	76	68
65	38319	Female	18	Unemployed	5	3	52	59	73	70
66	75891	Female	20	E1	6	4	54	63	76	71
67	11980	Female	18	Education	7	4	53	61	75	72
68	17762	Female	18	Unemployed	6	3	51	62	70	68
69	95621	Male	18	E2	6	3	53	59	74	71
70	66916	Male	18	E2	5	2	55	60	74	69
71	72150	Female	18	Education	3	2	55	61	72	70
72	64052	Female	20	E1	4	1	54	62	73	71
73	74267	Female	18	Education	5	3	55	61	73	71
74	52604	Female	18	Education	6	3	54	60	76	70
75	53467	Female	19	E2	5	1	51	60	74	68
76	91524	Male	18	Unemployed	6	4	54	59	75	68
77	92138	Female	19	Education	4	3	55	59	73	69
78	16245	Female	18	E2	7	5	54	57	74	69
79	85936	Female	21	E1	5	3	54	61	79	74
80	11607	Male	18	Education	5	2	54	60	76	67
81	97551	Female	18	E2	7	3	53	63	73	72
82	26252	Male	18	E1	5	4	54	63	74	74
83	68078	Female	19	E1	7	3	54	62	77	72
84	28774	Female	18	Unemployed	6	2	52	59	74	70
85	24121	Male	18	Unemployed	4	2	55	57	75	68
86	82826	Male	18	E1	6	4	54	63	74	73
87	72205	Male	19	E1	5	4	57	63	75	73
88	18130	Male	21	Unemployed	6	3	54	57	80	71
89	23682	Female	18	E2	5	1	53	58	75	66
90	52433	Female	20	E2	5	3	56	61	76	70
91	45790	Female	21	E2	7	4	55	62	75	71
92	80798	Female	21	E1	6	5	56	62	77	72
93	52926	Male	20	E2	6	3	56	60	76	68

94	85560	Female	20	Unemployed	5	2	53	60	75	70
95	29288	Female	18	Unemployed	6	3	54	61	76	68
96	98875	Male	20	E2	6	3	53	63	74	71
97	94873	Male	18	Education	7	4	53	62	76	72
98	86511	Male	21	Unemployed	7	2	50	61	69	68
99	69281	Male	18	E1	6	3	55	61	75	70
100	18891	Female	20	E1	5	3	56	62	75	72
101	55070	Female	19	Unemployed	6	4	57	61	75	69
102	19957	Female	18	Unemployed	6	5	55	60	75	71
103	74066	Female	18	Unemployed	5	3	55	59	75	71
104	76837	Female	21	Unemployed	6	2	51	59	71	66
105	59163	Male	19	Unemployed	6	2	52	57	74	67
106	28307	Male	18	Unemployed	6	2	49	56	71	67
107	61942	Female	18	E2	5	2	52	58	75	70
108	36050	Male	19	Education	6	3	55	61	76	71
109	40309	Male	21	E1	8	4	53	65	74	73
110	96729	Male	19	E1	6	4	57	62	76	72
111	53816	Male	18	E2	7	4	55	61	75	71
112	42624	Male	20	Education	6	2	55	58	77	71
113	17615	Male	19	Unemployed	6	3	54	61	76	72
114	80064	Male	20	Unemployed	3	1	54	53	76	67
115	74313	Female	18	E1	5	3	57	61	76	71
116	87662	Male	19	E1	6	5	56	63	73	72
117	20204	Female	18	Unemployed	5	3	51	59	74	68
118	13041	Female	18	Unemployed	6	3	52	58	72	68
119	26482	Female	18	E2	5	2	56	61	74	69
120	27745	Male	18	Unemployed	6	2	51	56	71	68
121	20889	Female	18	Education	6	2	54	63	74	69
122	82023	Female	18	E1	5	4	56	62	75	73
123	21272	Male	19	Education	7	2	54	63	75	70
124	53470	Male	18	Unemployed	5	3	51	57	72	68
125	15028	Male	18	E1	5	2	53	60	74	72
126	62725	Female	19	E2	7	4	55	60	77	69
127	46250	Male	20	Unemployed	5	3	53	61	71	68
128	54398	Male	18	E1	6	2	53	62	76	71
129	62240	Female	18	Unemployed	4	3	54	61	69	70
130	98981	Female	18	E2	6	3	54	61	76	70
131	50318	Female	18	Education	5	3	52	57	75	70
132	81994	Male	22	E2	7	3	56	63	73	70
133	84444	Male	18	Unemployed	5	2	53	61	72	70
134	83121	Male	21	Unemployed	6	3	54	62	73	72
135	87044	Female	18	Education	6	4	56	60	76	71
136	48626	Female	19	Education	6	4	52	61	72	72
137	99908	Female	18	Unemployed	4	2	55	60	72	69
138	56217	Male	23	Unemployed	6	3	53	59	76	70
139	37648	Female	18	E1	6	3	57	60	78	69
140	99589	Female	18	Education	5	4	53	60	73	72
141	45650	Female	18	E1	6	3	56	65	76	73
142	76951	Male	20	Unemployed	5	4	53	57	74	71
143	93460	Male	18	Unemployed	4	2	52	61	70	69

144	53630	Male	19	E1	5	2	55	60	75	72
145	80635	Female	18	Unemployed	6	1	53	60	72	65
146	95215	Male	18	E2	5	1	52	60	74	66
147	58976	Female	18	Unemployed	6	3	56	61	73	71
148	49230	Male	18	Unemployed	6	2	51	57	73	69
149	42456	Female	20	E2	5	2	51	58	74	69
150	21542	Female	18	Unemployed	5	3	54	59	71	68
151	94539	Female	19	E1	6	3	52	62	72	68
152	55375	Male	20	Unemployed	6	3	52	60	71	70
153	17059	Male	20	E1	7	4	54	61	78	74
154	78439	Female	21	E1	6	4	54	63	76	73
155	89185	Male	19	Unemployed	6	3	54	62	72	71
156	27439	Female	18	Unemployed	5	1	51	56	67	67
157	55946	Male	18	Unemployed	5	3	54	62	73	71
158	25925	Female	21	E2	6	3	52	60	74	70
159	74945	Male	18	Unemployed	4	2	56	62	74	70
160	64786	Female	18	Unemployed	7	4	54	61	76	71
161	90793	Male	20	E1	6	3	56	65	76	69
162	28101	Male	21	E1	6	3	54	62	78	74
163	71818	Male	18	Unemployed	7	4	54	58	75	68
164	28417	Male	18	Unemployed	7	4	55	57	77	70
165	51613	Male	18	E1	7	5	57	64	81	74
166	83183	Female	19	E2	6	3	54	63	72	72
167	85445	Female	18	Education	7	4	50	60	74	74
168	94778	Female	22	Unemployed	6	3	55	63	75	67
169	22838	Female	18	Unemployed	6	3	54	60	76	68
170	57517	Female	20	E1	7	4	55	63	76	71
171	35542	Female	18	E2	5	3	52	59	72	69
172	84523	Female	20	E1	7	3	52	63	75	74
173	89560	Female	18	E1	5	3	52	62	71	72
174	59531	Male	18	E1	6	3	54	63	73	73
175	14399	Female	20	E1	6	3	53	63	73	71
176	69834	Male	18	Unemployed	7	4	55	60	75	70
177	98136	Male	21	E1	6	3	51	59	73	70
178	79541	Male	18	E1	6	3	52	60	76	70
179	21959	Male	18	Unemployed	5	4	54	61	72	72
180	25341	Female	18	Unemployed	6	2	51	64	72	71
181	23588	Male	19	Education	7	3	53	61	71	66
182	93629	Female	18	Education	6	3	55	61	74	71
183	25505	Male	20	E1	6	2	54	58	76	67
184	57753	Female	18	E1	6	4	58	61	76	68
185	68689	Male	18	E1	6	2	54	59	75	70
186	54499	Female	19	Unemployed	4	3	56	61	74	72
187	13485	Male	18	Unemployed	7	4	54	57	76	69
188	64981	Male	37	E1	4	3	55	62	75	73
189	96347	Female	29	E1	5	1	51	62	72	71
190	86065	Male	18	Unemployed	7	2	52	56	75	67
191	22439	Female	18	Unemployed	6	4	55	61	75	70
192	45512	Female	27	Unemployed	6	3	55	58	74	68
193	71413	Male	43	E2	8	4	54	59	76	69

```
194     94034 Female  41         E1        7         4    54    61    76    72
195     75204 Female  18 Unemployed        6         2    50    59    71    69
196     46490   Male  18 Unemployed        5         1    51    59    72    68
197     91707   Male  18 Unemployed        6         4    50    58    72    70
198     35080   Male  18  Education        5         2    52    64    70    70
199     83579   Male  56         E2        6         4    53    64    75    71
200     40751   Male  21 Unemployed        5         3    53    62    71    69
201     41915 Female  33         E1        6         4    54    63    74    74
202     99302 Female  18         E1        5         3    52    61    71    72
203     13237   Male  18         E2        4         2    54    61    70    71
204     23244   Male  24         E1        6         4    54    62    79    73
205     64245   Male  56         E1        7         4    54    63    76    72
206     47565   Male  21         E2        6         1    52    60    74    68
207     99578   Male  43  Education        6         4    56    60    75    71
208     21185   Male  64         E1        3         3    54    64    72    75
209     76956   Male  31         E1        6         3    53    63    71    70
210     25452   Male  28         E2        6         3    51    57    74    71
211     43344   Male  18         E1        7         3    54    63    76    72
212     26846 Female  18         E1        4         3    56    62    76    72
213     76977 Female  29 Unemployed        6         4    54    60    75    71
214     49826   Male  18 Unemployed        6         3    55    61    77    69
215     87425 Female  18         E2        6         4    56    62    73    71
216     59929 Female  25         E1        6         2    55    61    78    71
217     31100   Male  18         E1        7         5    54    61    77    75
218     14105   Male  63         E1        6         4    54    64    74    75
219     62031 Female  18         E1        6         5    56    59    79    73
220     21839   Male  24  Education        5         3    52    61    74    71
221     96575 Female  52 Unemployed        8         4    54    60    77    68
222     32044 Female  18 Unemployed        6         3    52    59    73    68
223     58587   Male  50 Unemployed        4         2    51    57    69    69
224     28440   Male  18         E1        6         4    54    61    77    72
225     60234   Male  18         E2        6         3    56    60    75    70
226     47207   Male  43         E1        7         3    58    61    79    69
227     71218 Female  50 Unemployed        4         2    52    61    73    73
228     57465   Male  22 Unemployed        5         4    54    60    73    71
229     78411 Female  18         E1        5         3    55    58    77    70
230     24848   Male  18  Education        5         2    55    60    74    70
231     58931   Male  58         E1        8         6    58    62    77    73
232     55236 Female  51         E1        6         4    55    62    77    73
233     24966   Male  47         E2        6         3    51    60    73    70
```

```
cat("All checks passed successfully\n")
```

```
All checks passed successfully
```

Question 2

A colleague suggests the following research question,

“do students perform differently in their final year relative to their performance at the start?”

To assess this research question:

- create a new variable for the difference between Year 4 mark and Year 1 Mark.
- show and interpret a confidence interval for the mean difference, in context of the research question.

(12 marks)

Answer:

The results show that, on average, students’ performance in their final year is approximately 16.54 points higher than their performance at the start. The confidence interval provides additional information, suggesting that we can be reasonably confident that the true mean difference in performance between Year 4 and Year 1 lies between 16.25 and 16.84 points. This means that, based on the data, we can say with a certain level of confidence that the performance improvement in Year 4 relative to Year 1 is statistically significant and likely falls within this range. This suggests that there is evidence to conclude that students’ performance has improved over the years, as the confidence interval does not include zero.

In the context of the research question, these results indicate that there is a noticeable increase in students’ performance from the beginning to the final year of their study, and this increase is statistically supported by the confidence interval.

```
# Calculate the difference between Year 4 and Year 1
student_dataset$Year1vs4_difference <- student_dataset$Year4 -
student_dataset$Year1

# Calculate the mean difference
mean_difference <- mean(student_dataset$Year1vs4_difference)

# Calculate the confidence interval
confidence_interval <- t.test(student_dataset$Year1vs4_difference)$conf.int

# Print the results
cat("Mean Difference:", round(mean_difference, 2), "\n")

Mean Difference: 16.54

cat("Confidence Interval:", round(confidence_interval[1], 2), "to",
round(confidence_interval[2], 2), "\n")

Confidence Interval: 16.25 to 16.84
```

Question 3

Another research question is suggested,

"is there a relationship between student marks across each of the years?"

- Assess this research question by showing and interpreting the linear correlations between the marks for each of the four years.

Marks are awarded for well-designed output, and the interpretation of the output.

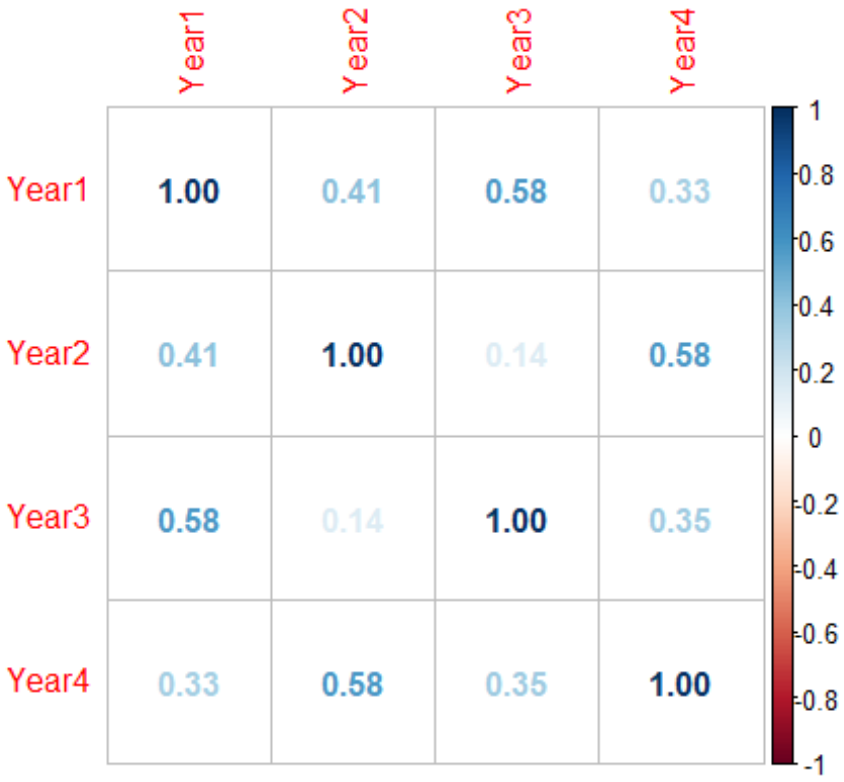
(12 marks)

Answer:

The correlation coefficients suggest that there are varying degrees of positive linear relationships between marks in different years. The relationships are moderate in some cases (Year 1 & Year 2, Year 1 & Year 3, Year 2 & Year 3) since the correlation coefficients are between 0.4 and 0.6, and weak or almost non-existent in others (Year 1 & Year 4, Year 2 & Year 4, Year 3 & Year 4) since the correlation coefficients are less than 0.4. Based on these results, we observed weaker correlations for Year 4 marks with all other years. Therefore, Year 4 appears to be less influenced by the academic performance of previous years compared to how Year 1, Year 2, and Year 3 are influenced. One possible interpretation is that students' marks in Year 4 are influenced by some external factors, such as their motivation based on their plan for the future or their attention to the job market, which causes them to focus less on lessons. In summary, the weaker correlations between Year 4 marks and those of previous years suggest that Year 4 performance is influenced by a unique set of factors or requirements, which may not be strongly tied to the linear progression of academic performance seen in earlier years. It highlights the distinct nature of Year 4 in the academic journey of students.

```r
# Calculate the correlation matrix
cor_matrix <- cor(student_dataset[, c("Year1", "Year2", "Year3", "Year4")])

# Create a correlation plot to show correlation values
library(corrplot)
corrplot(cor_matrix, method="number")
```



Question 4

A further research question states,

"can the final year mark be predicted based on one mark for a previous year?"

Produce simple linear regression with Year 4 mark as the dependent variable, and only one independent variable.

Your answer should include:

- justification for the choice of explanatory variable, including any additional supporting exploratory data analyses used to make the choice;
- interpretation of the slope (gradient) coefficient;
- comment on the r-square value, and the validity of model assumptions.

(23 marks)

Answer:

I chose Year 3 as the independent variable for predicting Year 4 marks because it not only had a relatively higher correlation with Year 4 marks (based on results for Question 3) but also provided a favourable goodness of fit. The AIC comparison confirmed that Year 2 also

had a strong prediction capability, but Year 3 remains relevant due to its contextual importance and moderate correlation.

The estimated slope coefficient (approximately 0.3362) means that for every one-point increase in a student's Year 3 marks, we can expect an increase of approximately 0.3362 points in their Year 4 marks, assuming a linear relationship between the two variables.

The R-squared value, which measures the proportion of variability in Year 4 marks explained by Year 3 marks, is approximately 0.1247 (or 12.47%). This suggests that only a small fraction of the variation in Year 4 marks can be explained by Year 3 marks alone. In other words, Year 3 marks are not a strong predictor of Year 4 marks, as indicated by the relatively low R-squared value.

The p-value is 3.203e-08 (much less than 0.05), indicating a statistically significant relationship between Year 3 marks and Year 4 marks. This means it's unlikely that this relationship is due to random chance alone.

In summary, while there is a statistically significant relationship between Year 3 marks and Year 4 marks, the strength of this relationship is relatively weak (as indicated by the low R-squared value). Therefore, predicting a student's Year 4 marks solely based on their Year 3 marks may not be very accurate or reliable. Other factors not included in this model may also influence Year 4 performance.

```
# Simple linear regression with Year 4 as the dependent variable
model_simple1 <- lm(Year4 ~ Year1, data = student_dataset)
model_simple2 <- lm(Year4 ~ Year2, data = student_dataset)
model_simple3 <- lm(Year4 ~ Year3, data = student_dataset)

# Compare models by AIC method
extractAIC(model_simple1)

[1]   2.0000 316.5255

extractAIC(model_simple2)

[1]   2.0000 249.4038

extractAIC(model_simple3)

[1]   2.0000 312.6895

# Based on correlation coefficients and AIC, I chose Year 3 as the
explanatory variable
summary(model_simple3)


Call:
lm(formula = Year4 ~ Year3, data = student_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-5.2135 -1.2135  0.1141  1.1227  5.4588

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.33610    4.35852  10.402  < 2e-16 ***
Year3        0.33618    0.05872   5.726  3.2e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.953 on 230 degrees of freedom
Multiple R-squared:  0.1247,    Adjusted R-squared:  0.1209
F-statistic: 32.78 on 1 and 230 DF,  p-value: 3.203e-08
```

```r
# Print the slope coefficient and R-squared value
slope <- coef(model_simple3)["Year3"]
cat("The slope (gradient) coefficient is:", round(slope, 4), "\n")
```

```
The slope (gradient) coefficient is: 0.3362
```

```r
r_squared <- summary(model_simple3)$r.squared
cat("The R-squared value is:", round(r_squared, 4), "\n")
```

```
The R-squared value is: 0.1247
```

Question 5: Report

Clearly state one **alternative new** research question based on the full original data set supplied to you. Explain why this is a worthwhile research question to consider.

You are required to write a short report for the client showing some analyses based only on the research question you have selected.

In your report you may wish to include a number of the following: exploratory data analyses; a hypothesis test; data modelling; discussion of limitations; how you could extend the research if given more time.

To clarify, your answer to this question must be a report based on **your analyses of your own research question** arising from the data, which is not addressed in the questions above. This report should contain a maximum 5 outputs (i.e. graphics + tables) and a maximum of 500 words.

(35 marks)

Answer:

This research study aims to explore the factors that influence student performance, a matter of great significance for educational institutions and stakeholders. Specifically, I investigate the impact of students' language proficiency, measured by the "language" variable, on their academic performance, represented by the "mark" variable. To facilitate

this analysis, a new variable, “Mark,” was created by averaging the values from four “Year” columns. Understanding the influence of language skills on academic outcomes is vital, as language proficiency is fundamental for effective communication and comprehension. Additionally, I examine the relationship between students’ performance and their educational outcomes.

The hypothesis testing results indicated a statistically significant positive correlation between language proficiency and academic performance. The simple linear regression model further supported this finding, with a slope coefficient of 0.3023 and a small p-value, suggesting that, on average, a one-unit increase in “language” is associated with a 0.3023 unit increase in estimated “Mark.” However, it’s essential to note that the model’s R-squared value suggests that “language” explains only approximately 4.33% of the variance in “Mark.” This implies that other unaccounted factors may also influence students’ performance.

For a deeper analysis, a multiple linear regression model was constructed, with both “language” and “age” as independent variables. While the R-squared value saw a slight increase, the p-value for the “age” variable (0.02151) indicated significance but contributed modestly. The stepAIC model confirmed the importance of “language” by showing significant changes in AIC values upon its removal. Thus, “language” plays a crucial role in the model and should be retained as a significant predictor.

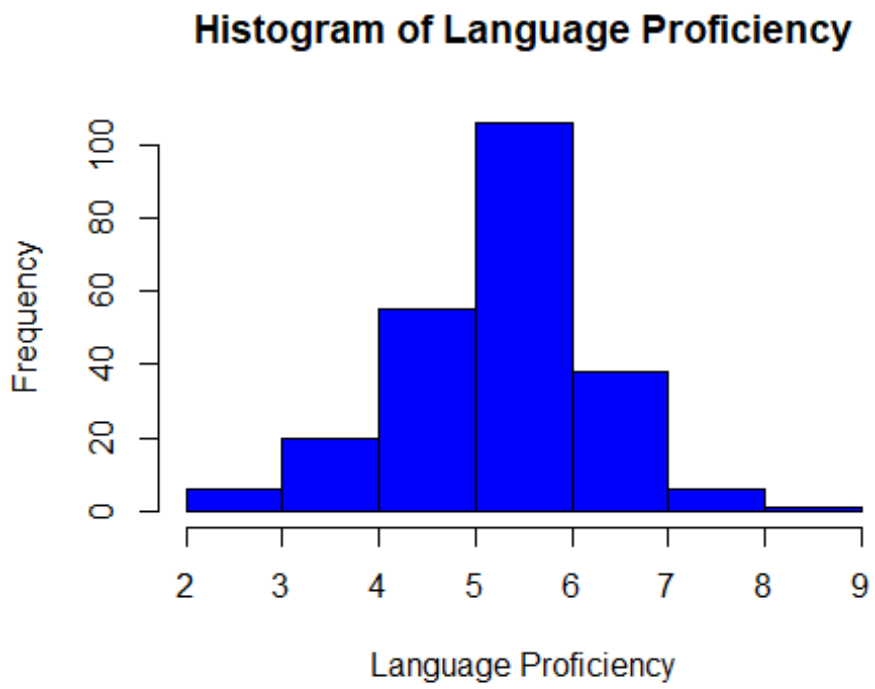
Visualisations, such as scatter plots, supported the findings by demonstrating that students with stronger language skills tend to perform better academically, as their average marks are higher. Histograms provide a visual representation of the distribution of “language” and “mark” variables, helping us understand their concentration. Additionally, box plots revealed that students with higher marks and superior academic performance are more likely to be employed in graduate roles, highlighting the potential career benefits associated with higher academic achievement. This finding underscores the significance of academic performance in shaping students’ future prospects. Therefore, it is essential to continue investigating the factors that enhance academic performance.

While this report covers essential analyses, further exploratory data analyses could uncover additional insights. It is essential to recognise the study’s limitations. For instance, there may be unmeasured variables that impact student performance that were not considered in this analysis. Additionally, the study’s scope is limited to the available variables, and other unrecorded factors may contribute to variations in student performance. Moreover, it is important to note that linear regression assumes certain conditions like normality and homoscedasticity, which should be considered when interpreting the results.

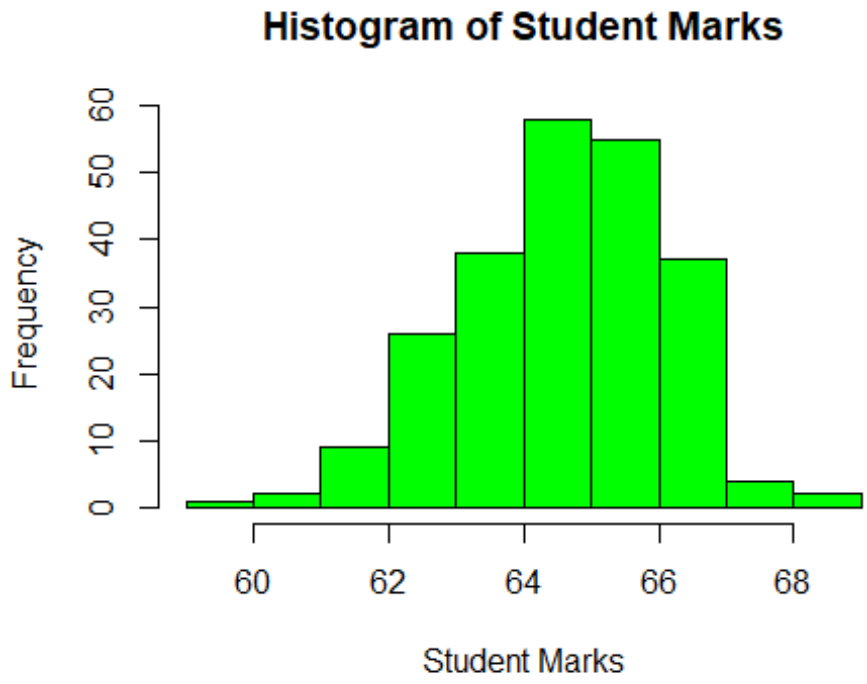
```
# Research question: "Does language proficiency impact academic performance
and consequently future outcomes in students?"

# Creating a new variable "Mark" by calculating the average of the four
"Year" variables.
student_dataset$Mark <- rowMeans(student_dataset[, c("Year1", "Year2",
"Year3", "Year4")])
```

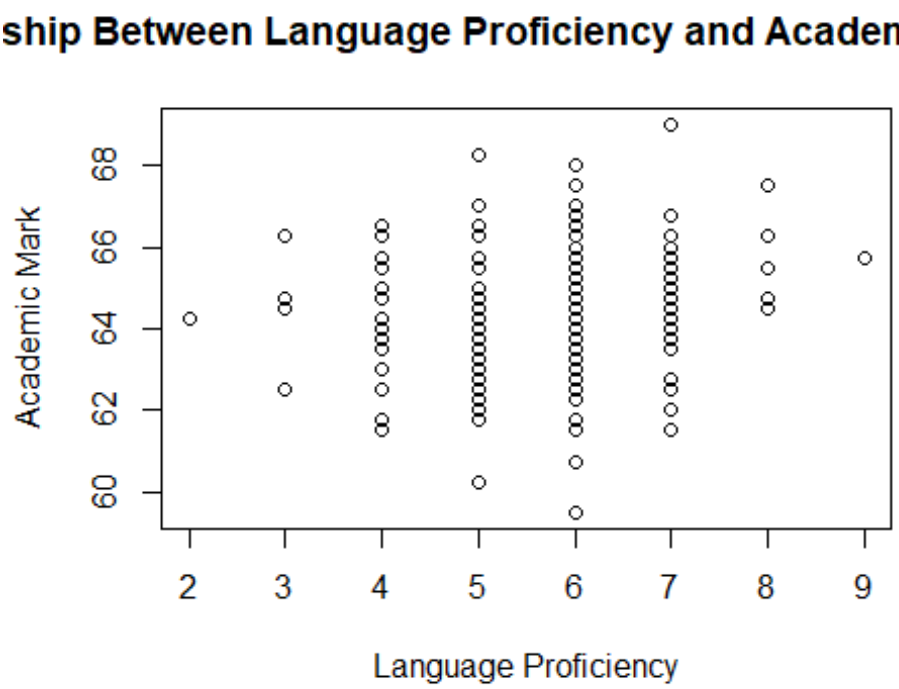
```r
# Exploring distribution of language proficiency and Mark
hist(student_dataset$language, main = "Histogram of Language Proficiency",
xlab = "Language Proficiency", col = "blue")
```




```r
hist(student_dataset$Mark, main = "Histogram of Student Marks", xlab =
"Student Marks", col = "green")
```




```r
# Visualize the relationship between language proficiency and academic performance by scatter plot
plot(student_dataset$language, student_dataset$Mark, xlab = "Language Proficiency", ylab = "Academic Mark", main = "Relationship Between Language Proficiency and Academic Performance")
```




```r
# Hypothesis Test
corr_lang_mark <- cor(student_dataset$language, student_dataset$Mark)
cat("Correlation between language and Mark:", corr_lang_mark, "\n")
```

```
Correlation between language and Mark: 0.2079707
```

```r
p_value <- cor.test(student_dataset$language, student_dataset$Mark)$p.value
if (p_value < 0.05) {
  cat("The p-value is", p_value, "which is less than 0.05. Therefore, we
reject the null hypothesis.")
  cat("There is a statistically significant positive correlation between
language proficiency and academic performance.")
} else {
  cat("The p-value is", p_value, "which is greater than 0.05. Therefore, we
fail to reject the null hypothesis.")
  cat("There is no statistically significant correlation between language
proficiency and academic performance.")
}
```

```
The p-value is 0.001445003 which is less than 0.05. Therefore, we reject the
null hypothesis.There is a statistically significant positive correlation
between language proficiency and academic performance.
```

```r
# Linear Regression Model
model_simple <- lm(Mark ~ language, data = student_dataset)
summary(model_simple)
```

```
Call:
lm(formula = Mark ~ language, data = student_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2864 -0.7995  0.1613  0.9636  3.9113

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 62.97246    0.54665 115.196  < 2e-16 ***
language     0.30232    0.09376   3.225  0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.485 on 230 degrees of freedom
Multiple R-squared:  0.04325,   Adjusted R-squared:  0.03909
F-statistic:  10.4 on 1 and 230 DF,  p-value: 0.001445
```

```
# Considering other factor (age) in the model
model_multiple <- lm(Mark ~ language + age, data = student_dataset)
summary(model_multiple)
```

```
Call:
lm(formula = Mark ~ language + age, data = student_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2236 -0.8288  0.1240  0.9660  4.0164

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 62.47933    0.58194 107.363  < 2e-16 ***
language     0.28735    0.09311   3.086  0.00228 **
age          0.02737    0.01183   2.315  0.02151 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.471 on 229 degrees of freedom
Multiple R-squared:  0.06513,   Adjusted R-squared:  0.05696
F-statistic: 7.977 on 2 and 229 DF,  p-value: 0.0004479
```

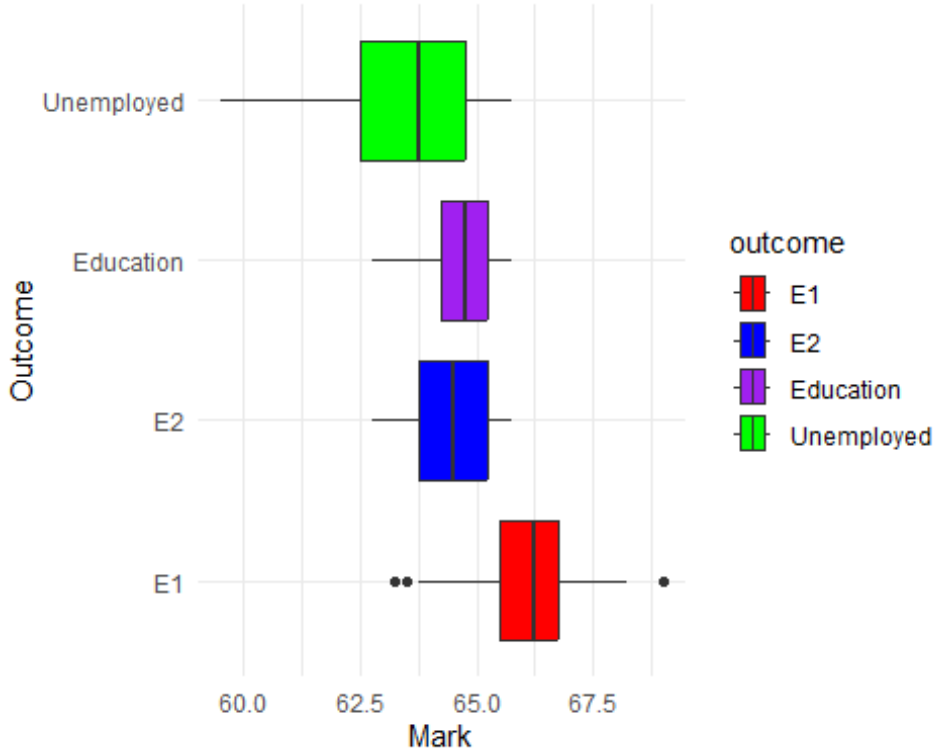
```
model_final <- stepAIC(model_multiple)
```

```
Start:  AIC=182.18
Mark ~ language + age

           Df Sum of Sq    RSS    AIC
<none>                  495.78 182.18
```

```
- age        1    11.601 507.38 185.55
- language   1    20.623 516.40 189.63

# Exploring the relationship between the academic performance and outcome
after graduation by boxplot
library(ggplot2)
ggplot(student_dataset, aes(x = Mark, y = outcome, fill = outcome)) +
  geom_boxplot() +
  labs(x = "Mark", y = "Outcome") +
  scale_fill_manual(values = c("E1" = "red", "E2" = "blue", "Unemployed" =
"green", "Education" = "purple")) +
  theme_minimal()
```



Question 6

Describe how you have applied principles of reproducible research in this submission (maximum 100 words).

Marks are awarded for identification of appropriate reproducible research principles, only if also evidenced throughout your submission that they have been applied.

(8 marks)

Answer:

Several principles of reproducible research have been applied to ensure that it can be easily reproduced and verified by others:

Extensive comments and explanations have been provided. This helps others understand the code's logic and purpose. Variables have been named descriptively, making it clear what each variable represents. Data cleaning and manipulation steps are clearly explained, ensuring that the finer dataset is well-defined and reproducible.

Results and interpretations have been presented in a clear and organised report format. Visualisation is used to present the results graphically, making it easier for others to interpret the findings. The report includes clear conclusions and interpretations based on the analysis.

End matter - Session Information

Do not edit this part. Make sure that you compile your document so that the information about your session (including software / package versions) is included in your submission.

```
sessionInfo()

R version 4.4.0 (2024-04-24 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 10 x64 (build 19045)

Matrix products: default


locale:
[1] LC_COLLATE=English_United States.utf8
[2] LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8

time zone: Europe/London
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] boot_1.3-30        ggplot2_3.5.1      dplyr_1.1.4        readr_2.1.5
 [5] flextable_0.9.6    psych_2.4.3        tableone_0.13.2
performance_0.11.0
 [9] corrplot_0.92      RSQLite_2.3.6      assertr_3.0.1      scales_1.3.0
[13] gapminder_1.0.0    MASS_7.3-60.2
```

```
loaded via a namespace (and not attached):
 [1] tidyselect_1.2.1        farver_2.1.2            blob_1.2.4
 [4] fastmap_1.2.0           fontquiver_0.2.1        promises_1.3.0
 [7] digest_0.6.35           mime_0.12               lifecycle_1.0.4
[10] gfonts_0.2.0            survival_3.5-8          magrittr_2.0.3
[13] compiler_4.4.0          rlang_1.1.3             tools_4.4.0
[16] utf8_1.2.4              yaml_2.3.8              data.table_1.15.4
[19] knitr_1.46              askpass_1.2.0           labeling_0.4.3
[22] bit_4.0.5               mnormt_2.1.1            curl_5.2.1
[25] xml2_1.3.6              httpcode_0.3.0          withr_3.0.0
[28] grid_4.4.0              fansi_1.0.6             gdtools_0.3.7
[31] xtable_1.8-4            colorspace_2.1-0        crul_1.4.2
[34] insight_0.19.11         cli_3.6.2               survey_4.4-2
[37] rmarkdown_2.27          crayon_1.5.2            ragg_1.3.2
[40] generics_0.1.3          rstudioapi_0.16.0       tzdb_0.4.0
[43] DBI_1.2.2               cachem_1.1.0            splines_4.4.0
[46] parallel_4.4.0          mitools_2.4             vctrs_0.6.5
[49] Matrix_1.7-0            jsonlite_1.8.8          fontBitstreamVera_0.1.1
[52] hms_1.1.3               bit64_4.0.5             systemfonts_1.1.0
[55] glue_1.7.0              gtable_0.3.5            later_1.3.2
[58] munsell_0.5.1           tibble_3.2.1            pillar_1.9.0
[61] htmltools_0.5.8.1       openssl_2.2.0           R6_2.5.1
[64] textshaping_0.3.7       evaluate_0.23           shiny_1.8.1.1
[67] lattice_0.22-6          highr_0.10              memoise_2.0.1
[70] fontLiberation_0.1.0    httpuv_1.6.15           Rcpp_1.0.12
[73] zip_2.3.1               uuid_1.2-0              nlme_3.1-164
[76] officer_0.6.6           xfun_0.44               pkgconfig_2.0.3
```