

Reading Habits

Final Project for Python Course, Data Science and Application -
Advanced Diploma, Metro College of Technology

August, 2021



Introduction

The importance of Reading is mentioned by all mentors and tutors since a very young age, but what is our reading habits nowadays?

This project aims to deepen this question to gather insights into reading habits by applying Exploratory Data Analysis.

Methodology

- Univariate analysis (plotting and independence tests)
- Bivariate analysis (plotting and independence tests)

Data source: <https://www.kaggle.com/vipulgote4/reading-habit-dataset>

“I read for pleasure and
that is the moment I learn
the most”

Margaret Atwood

Dataset

The original dataset has 14 features and 2,823 observations.

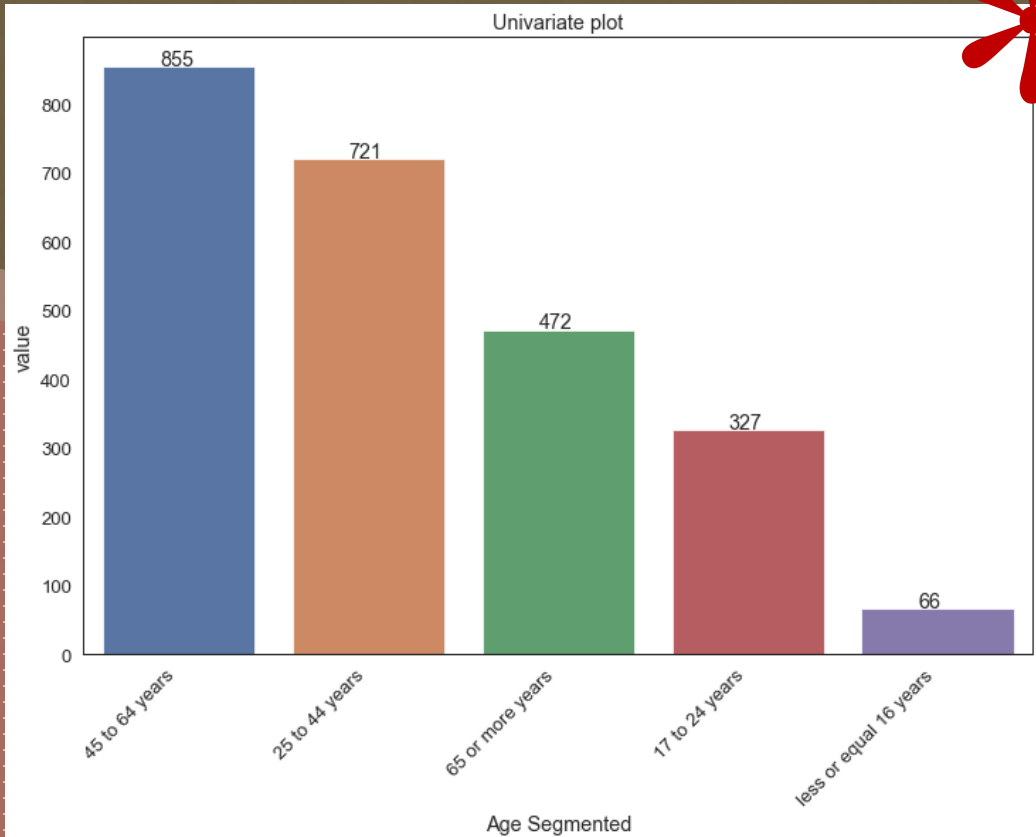
Feature	Type
Age	int64
Sex	object
Race	object
Marital status?	object
Education	object
Employement	object
Incomes	object
How many books did you read during last 12months?	int64
Read any printed books during last 12months?	object
Read any audiobooks during last 12months?	object
Read any e-books during last 12months?	object
Last book you read, you...	object
Do you happen to read any daily news or newspapers?	object
Do you happen to read any magazines or journals?	object

The Missing values were dropped as one duplicated value, remaining 2441 observations.

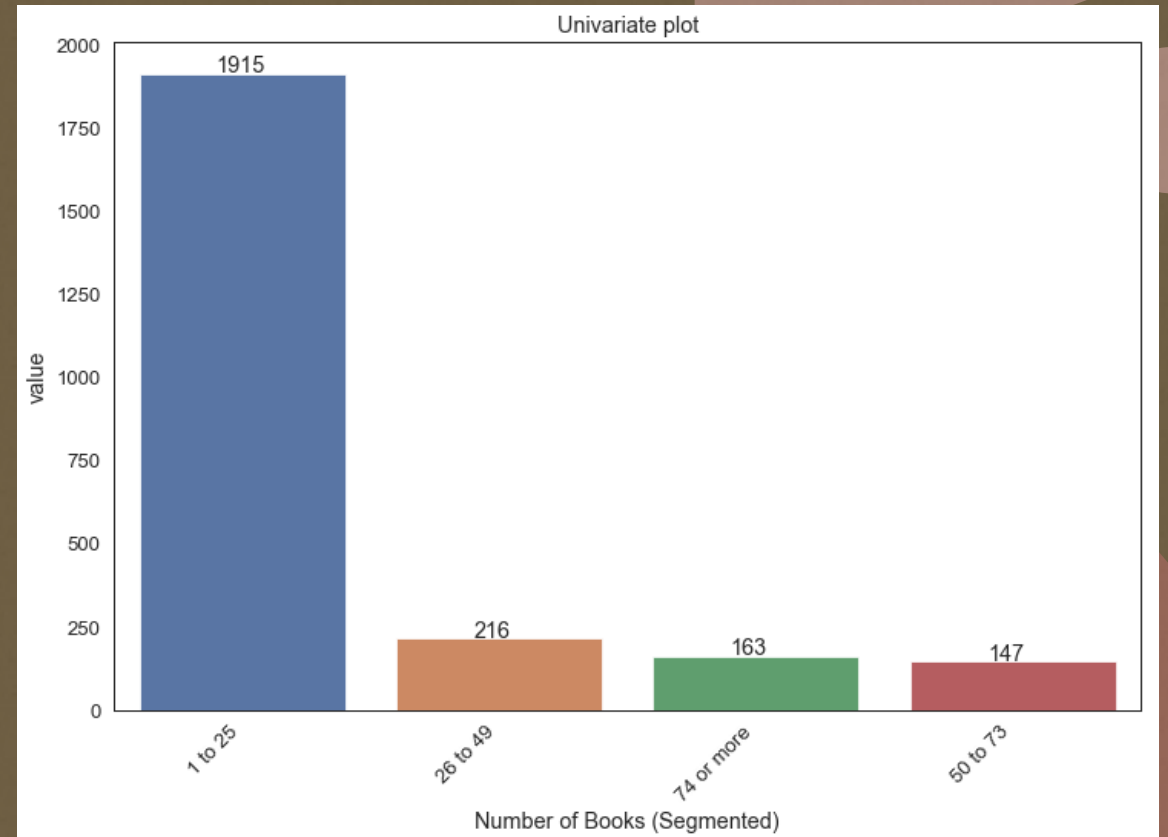
Features	Total	# Missing	% Missing
Age	2831	0	0
Do you happen to read any daily news or newspaper...	2831	0	0
Do you happen to read any magazines or journals?	2831	0	0
Education	2831	0	0
Employement	2831	0	0
How many books did you read during last 12months?	2831	0	0
Incomes	2831	0	0
Last book you read, you...	2831	390	13.78
Marital status?	2831	0	0
Race	2831	0	0
Read any audiobooks during last 12months?	2831	390	13.78
Read any e-books during last 12months?	2831	390	13.78
Read any printed books during last 12months?	2831	390	13.78
Sex	2831	0	0

Columns added

- Age Segmented:
 - less or equal 16 years
 - 17 to 24 years
 - 25 to 44 years
 - 45 to 64 years
 - 65 or more years

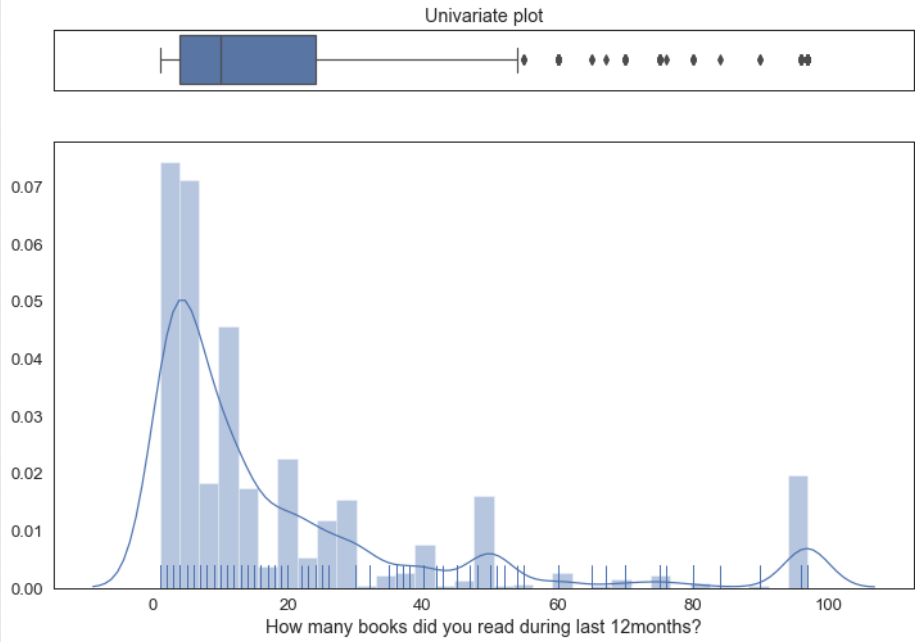
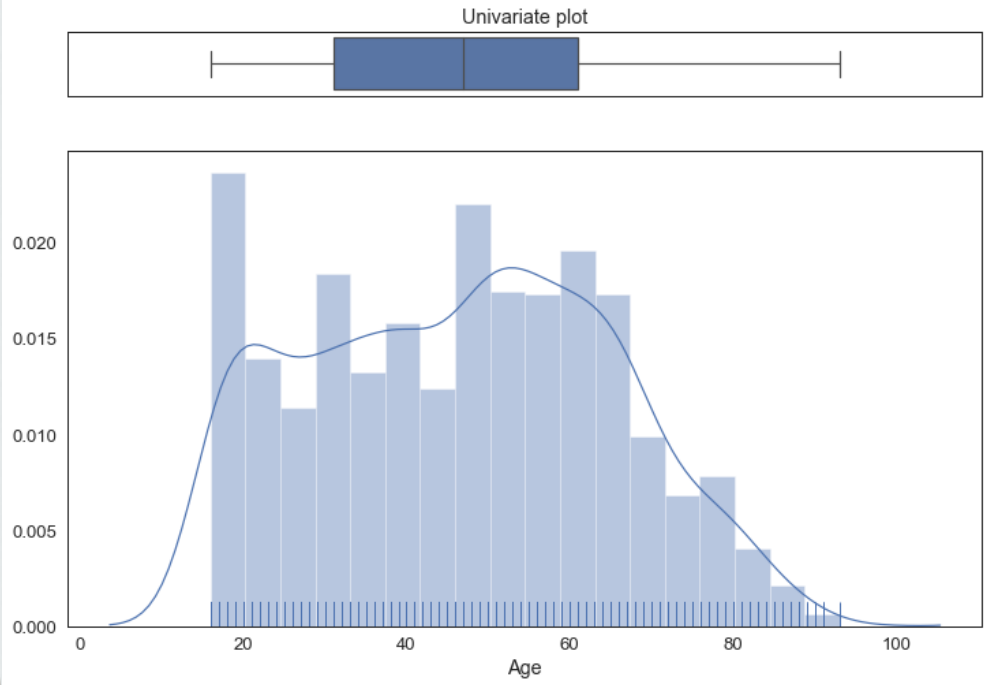


- Number of Books (Segmented):
 - 1 to 25
 - 26 to 49
 - 50 to 73
 - 74 or more



What are the ages and number of books distributions read in this survey?

	Age	How many books did you read during last 12months?
count	2441	2441
mean	46.5207	19.3425
std	18.4882	24.4212
min	16	1
25%	31	4
50%	47	10
75%	61	24
max	93	97
variance	341.812	596.397
IQR	30	20
range	77	96
skewness	0.070687	2.04029
kurtosis	-0.93216	3.52504
mode	16	3

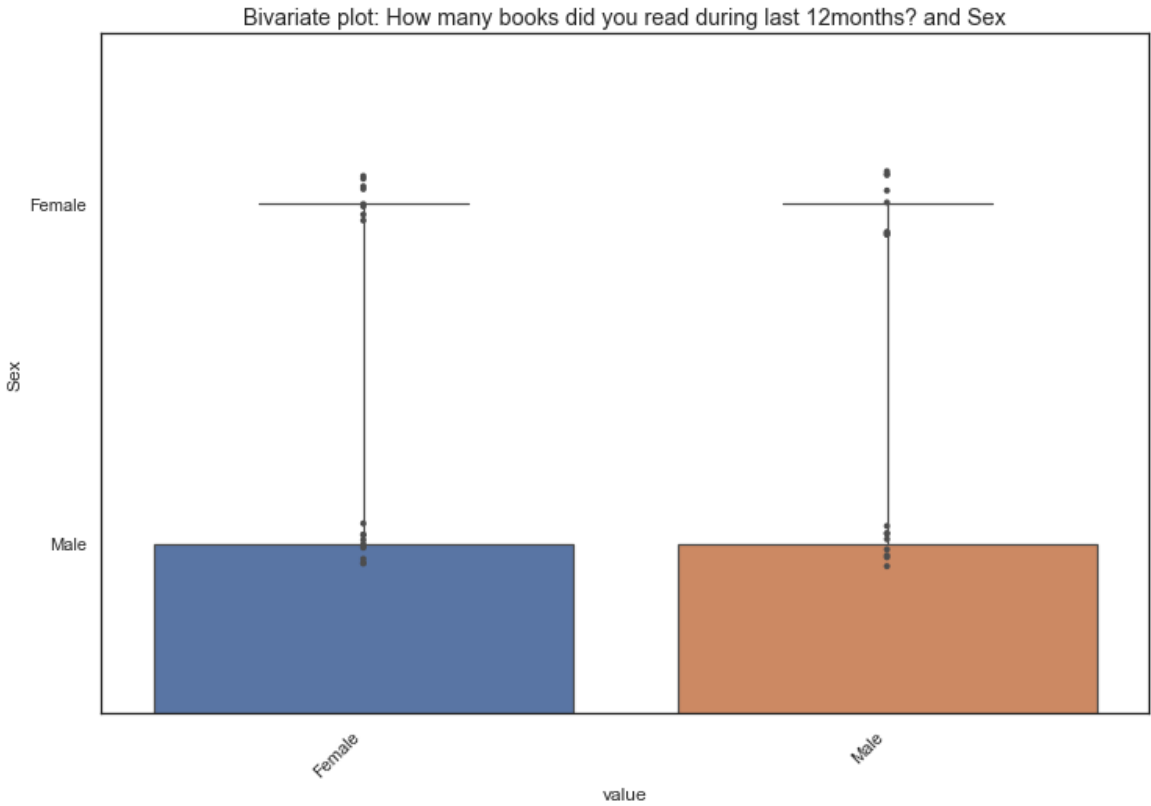
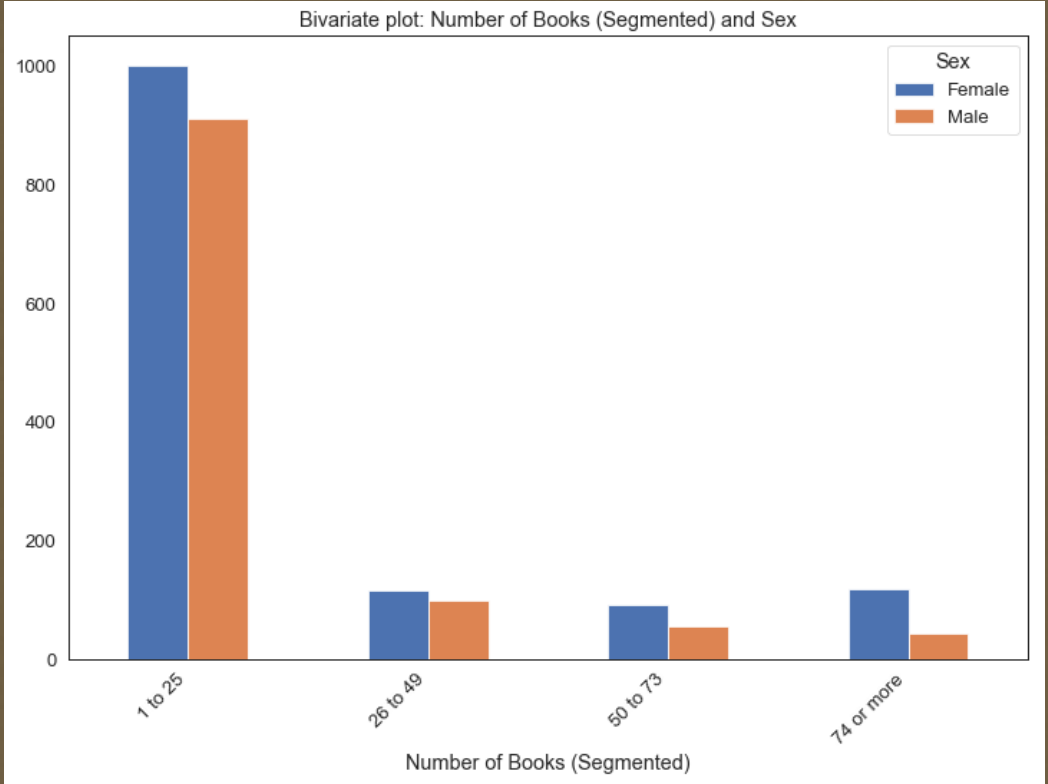


Which Gender reads the most?

Sex	N	Mean	Median	SD	SE	95% Conf. Interval
Female	1329	21.917	10	26.819	0.736	20.474 23.36
Male	1112	16.265	8	20.801	0.624	15.041 17.489

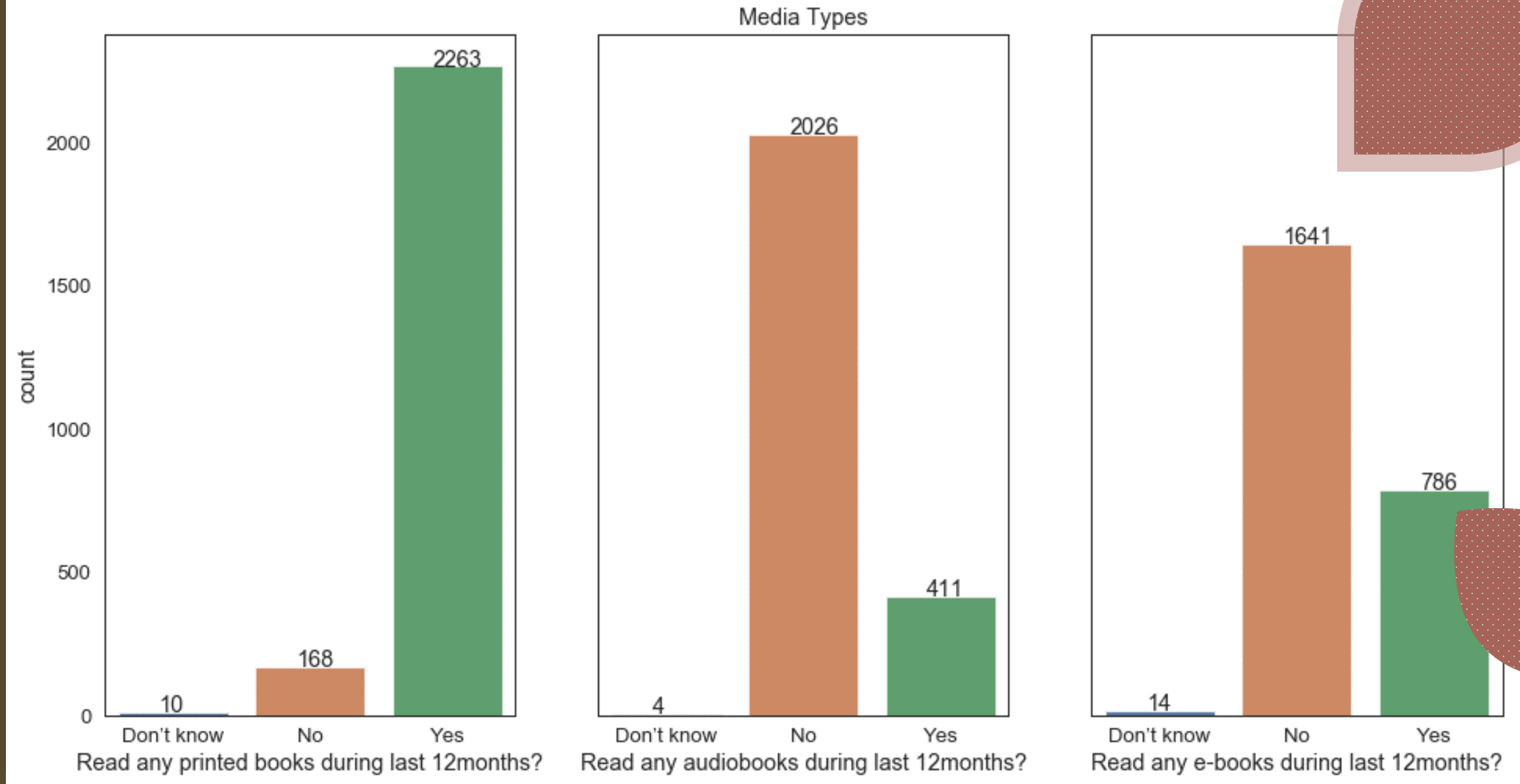
p-value is: p=0.000

Chi-square's test result: Features are dependent (reject H0)



Mann-Whitney U 's Statistics (unequal variances)=654059.500, p=0.000
Mann-Whitney U 's Null hypothesis: The distribution of scores for the two groups are equal.
Mann-Whitney U 's test result: The distribution of scores for the two groups are not equal. (reject H0)

What are the medias being used?



Is Age correlated to How many books one read in a year?

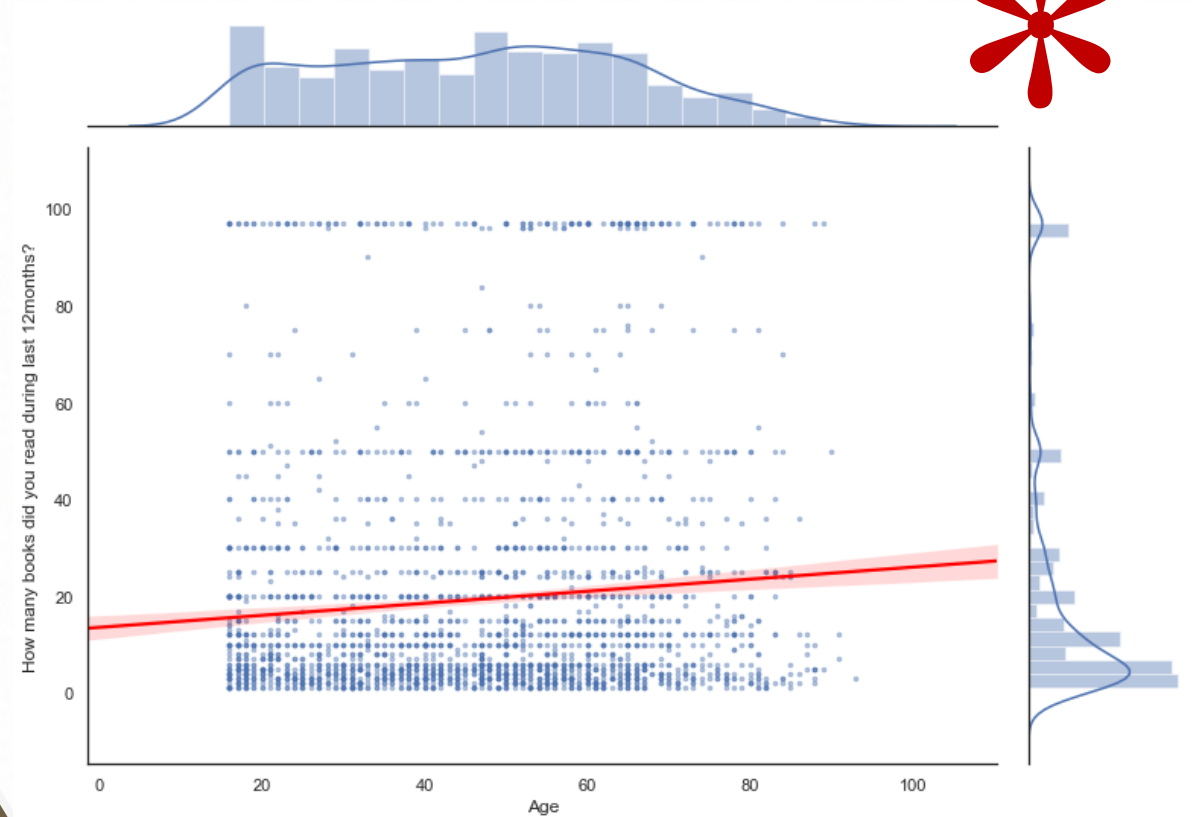
Age x How many books did you read during last 12months?

Pearson's correlation:

0.09417

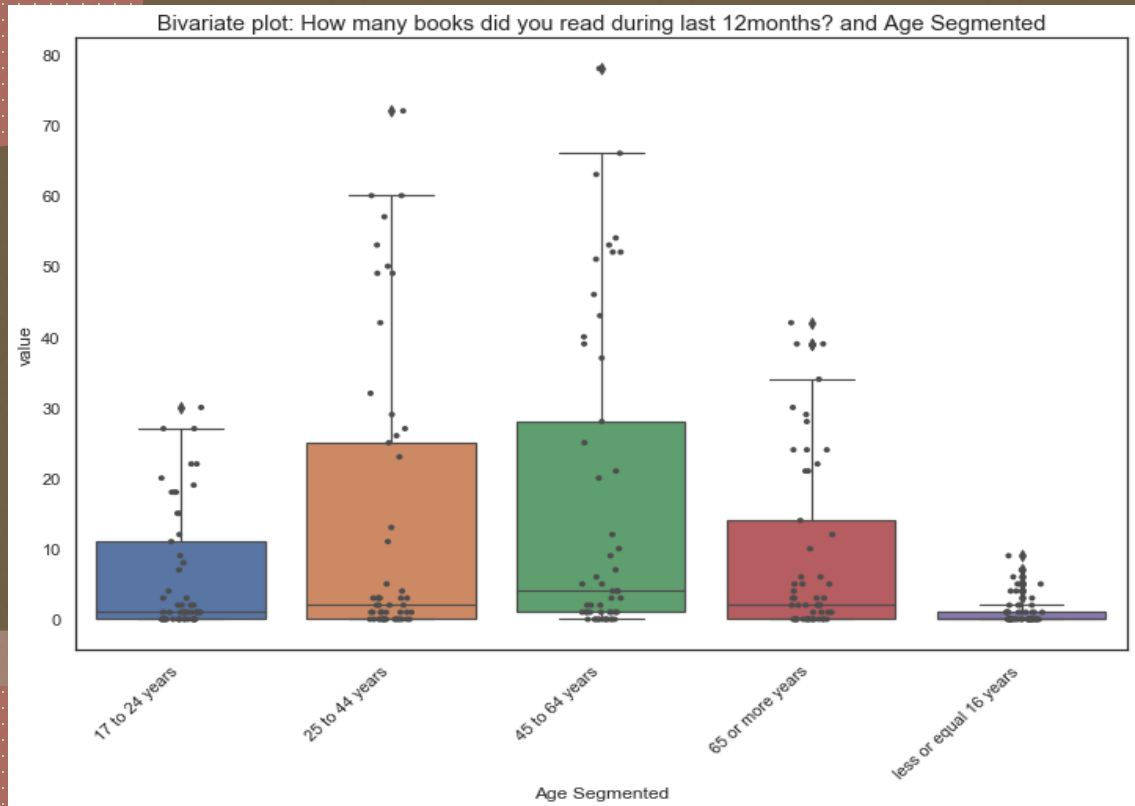
Spearman's correlation:

0.082017



*Age and number of books read were segmented to test this relation

Continue...



Chi-square p-value is: $p=0.000$

Chi-square's test result: Features are dependent (reject H_0)

Null hypothesis: Features are independent

Assumptions are not met in some of the number of observations individual item.

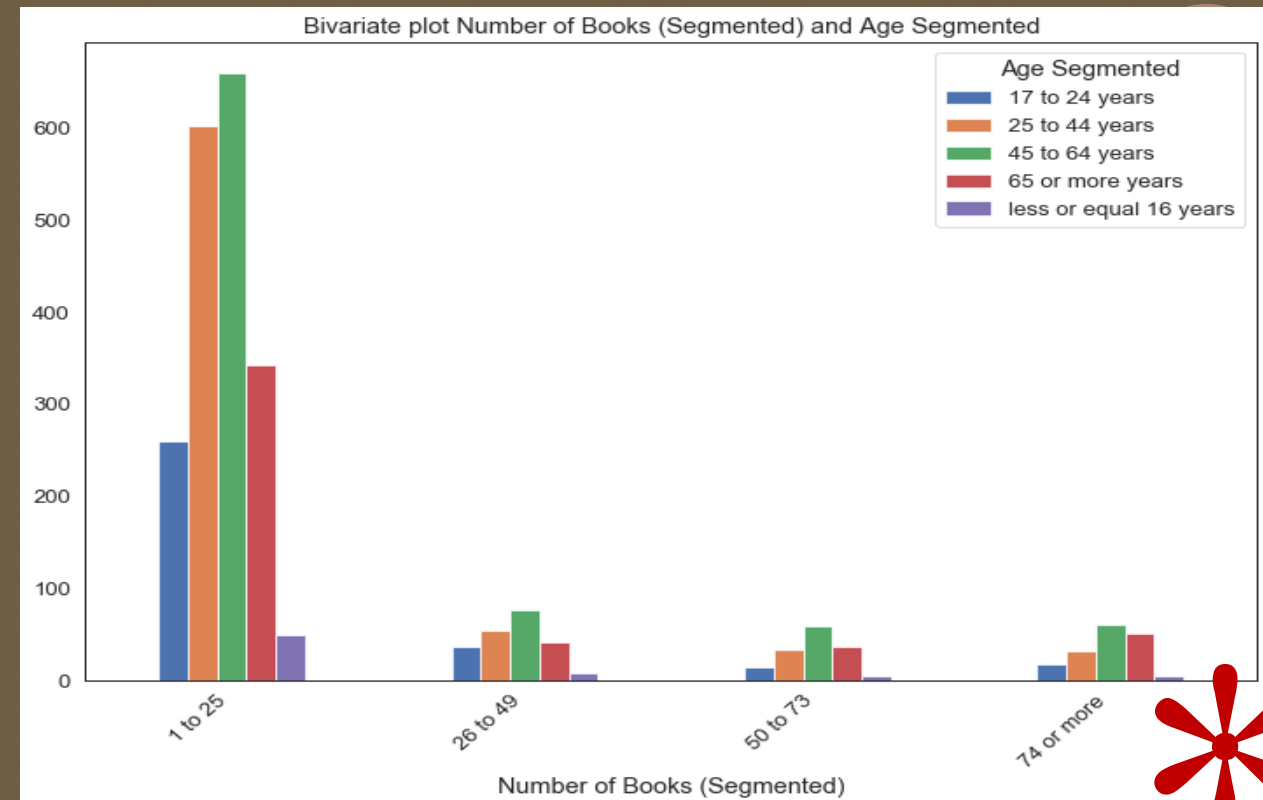
Chi-square item may not be reliable. Gather more data.

Kruskal's Statistics(unequal variances)=26.677, $p=0.000$

Kruskal's Null hypothesis: All sample distributions are equal.

Kruskal's test result: One or more sample distributions are not equal. (reject H_0)

*Independence tests point that there may be a relationship. But more data is necessary to confirm.



Does Education impacts reading habit?



*Independence tests point that there may be a relationship. But more data is necessary to confirm.

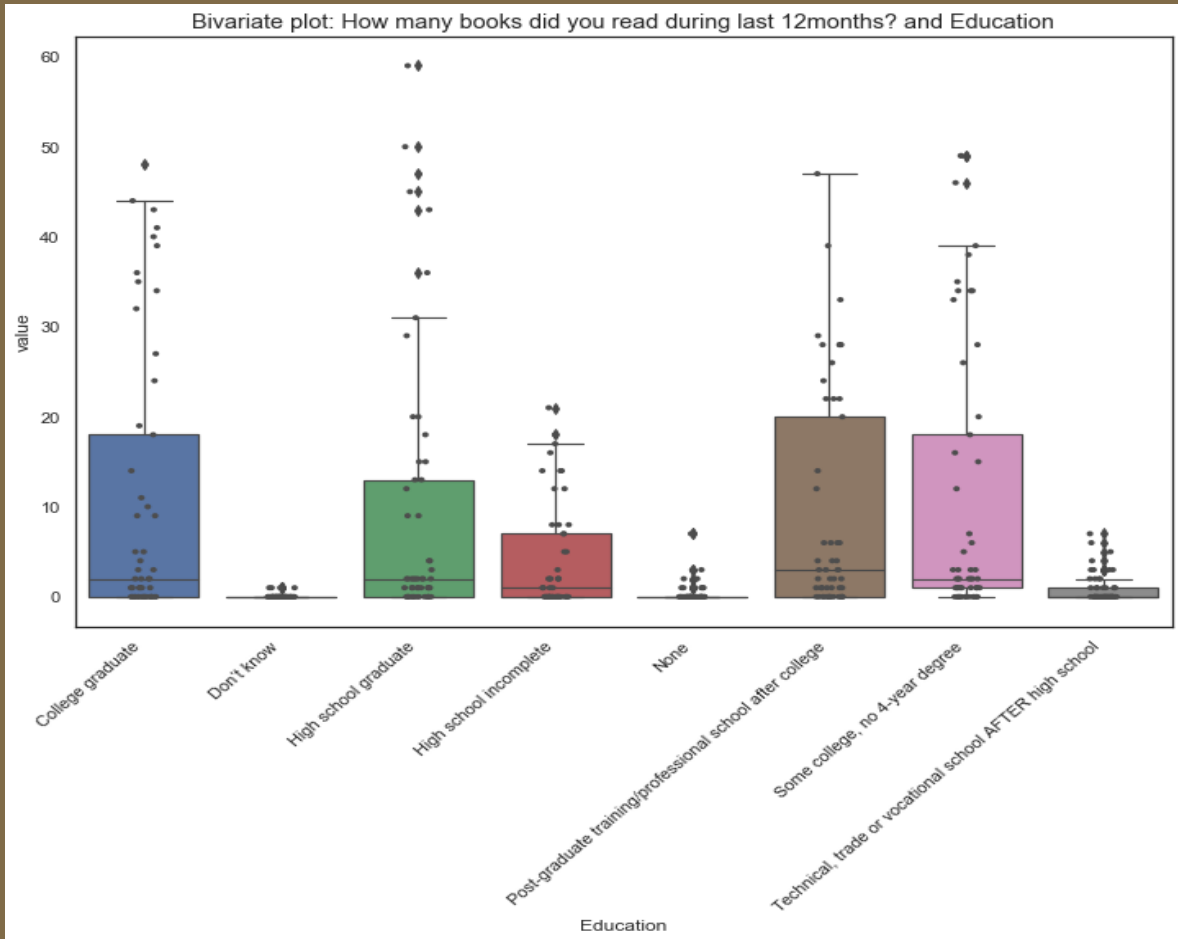
Chi-square p-value is: $p=0.000$

Chi-square's test result: Features are dependent (reject H_0)

Null hypothesis: Features are independent

Assumptions are not met in some of the number of observations individual item.

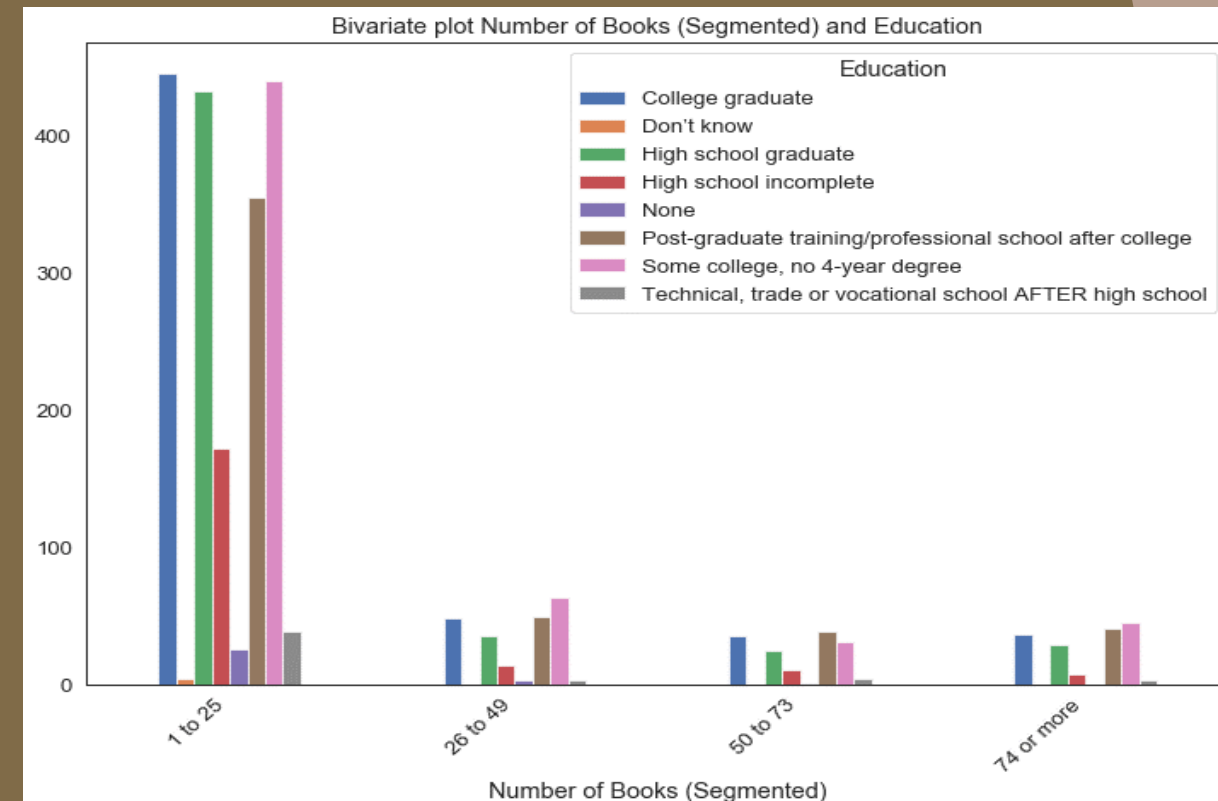
Chi-square item may not be reliable. Gather more data.



Kruskal's Statistics(unequal variances)=78.550, $p=0.000$

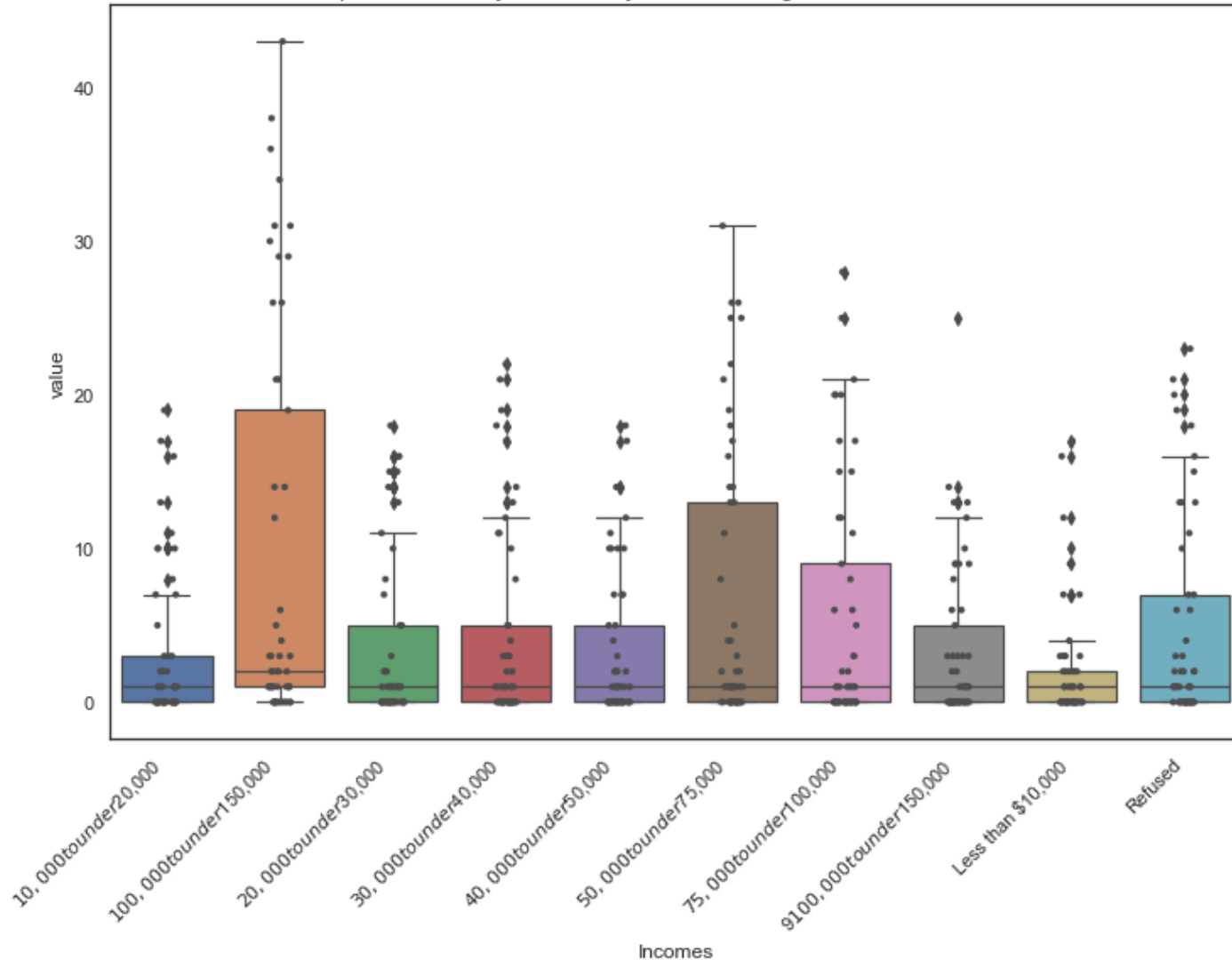
Kruskal's Null hypothesis: All sample distributions are equal.

Kruskal's test result: One or more sample distributions are not equal. (reject H_0)



Does Income impacts reading habit?

Bivariate plot: How many books did you read during last 12months? and Incomes



Kruskal's Statistics(equal variances)=31.141,
 $p=0.000$

Kruskal's Null hypothesis: All sample distributions
are equal.

Kruskal's test result: One or more sample
distributions are not equal. (reject H_0)

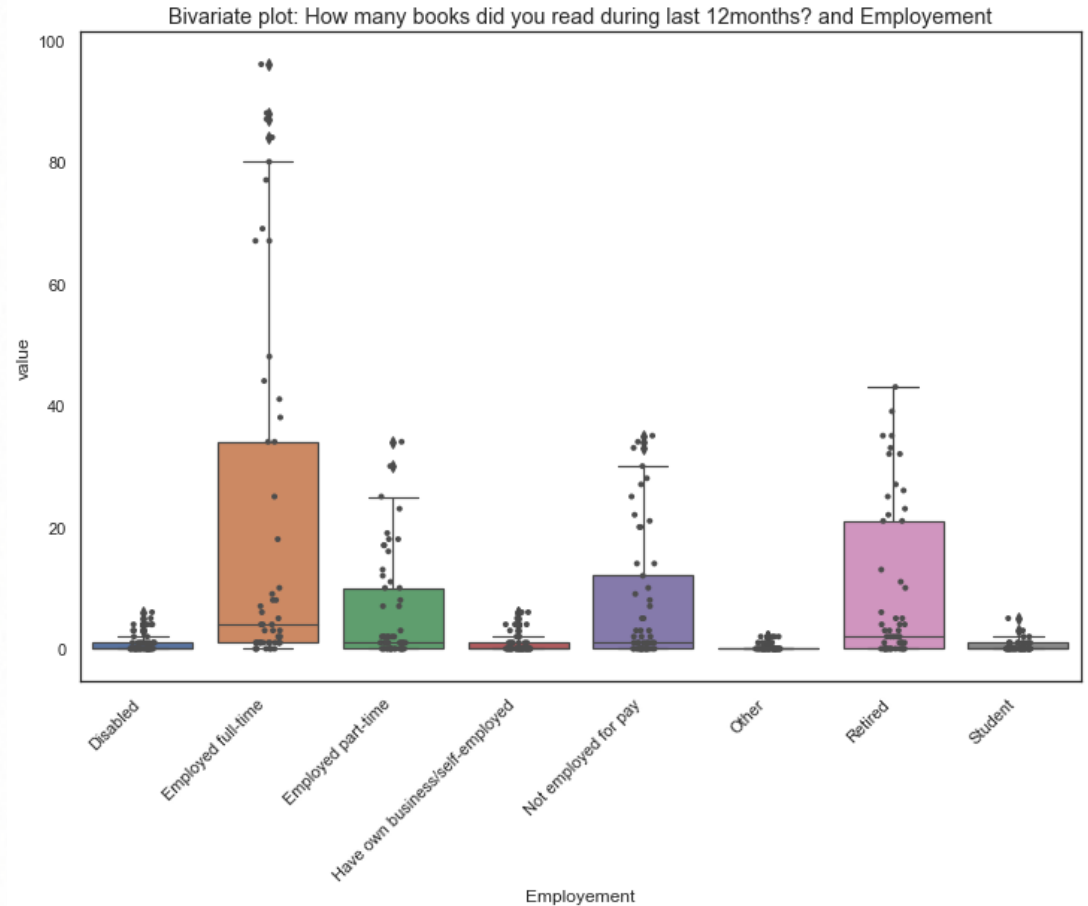
Does Employment status impacts reading habit?

Kruskal's Statistics(unequal variances)=23.078,
 $p=0.002$

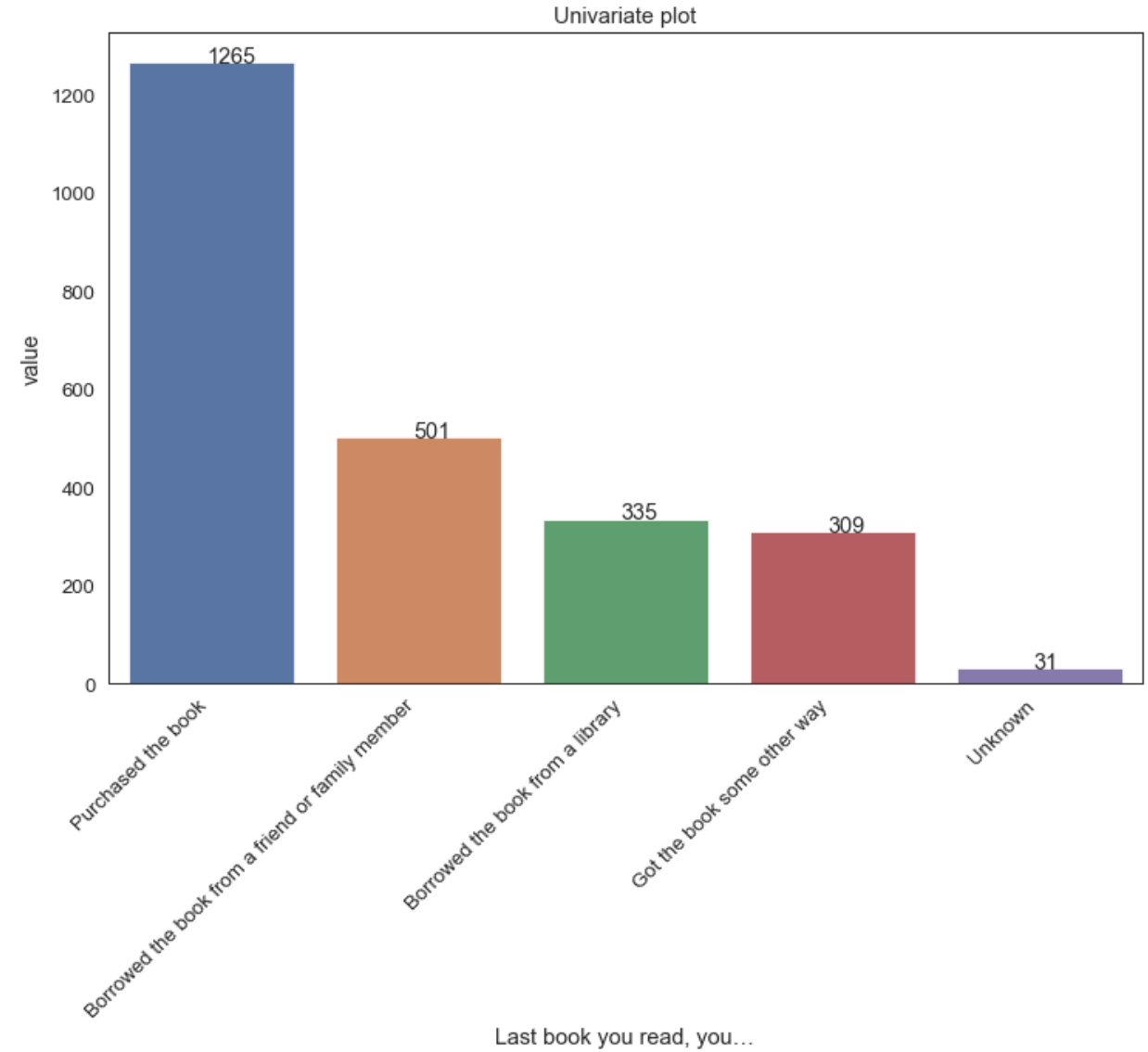
Kruskal's Null hypothesis: All sample
distributions are equal.

Kruskal's test result: One or more sample
distributions are not equal. (reject H_0)

Employment	N	Mean	SD	SE	95% Conf. Interval
Disabled	46	16.413	18.585	2.74	10.894 21.932
Employed full-time	1089	17.338	22.039	0.668	16.028 18.648
Employed part-time	322	19.537	24.606	1.371	16.839 22.235
Have own business/self-employed	47	21.894	31.444	4.587	12.661 31.126
Not employed for pay	402	18.152	23.423	1.168	15.855 20.448
Other	13	26.692	34.311	9.516	5.959 47.426
Retired	500	24.532	28.618	1.28	22.017 27.047
Student	22	15.864	20.081	4.281	6.96 24.767



How the last book was acquired?



Next steps...

- Gather more data
- Perform post-hoc tests to find which subgroups diverge in the group
- Study features trend through the years

Conclusions

- Females read more than males
- Word of mouth appears to be the second-best manner to rouse reading habit
- Printed books continue to be the media of choice
- Employment status is related to reading habit
- Despite the need to investigate further:
 - Better income distribution may be one key point to encourage reading habit (related to employment situation?)
 - Higher education appears to impact reading habits
- Age seems to be related to reading habit, but more data is necessary to confirm this hypothesis