# Reading Habits

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

# August,2021

Ana Clara Tupinambá Freitas, course ministered by Professor Tareq Jaber

# 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.



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

Margaret Atwood



# Methodology

• Business understanding;

• EDA:

– Univariate analysis; and

– Bivariate analysis.

# Dataset description

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

The data has **14 features** and **2,823 observations**. The features are:

|  |  |
| --- | --- |
| 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 |

# Initial analysis and actions

## Duplicated values:

Dropped, kept last observations.

## Missing values:

|  |  |  |  |
| --- | --- | --- | --- |
| Features | Total | # Missing | % Missing |
| Age | 2831 | 0 | 0 |
| Do you happen to read any daily news or newspap... | 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 |

Missing values were dropped, remaining **2441 observations**.

## Columns added:

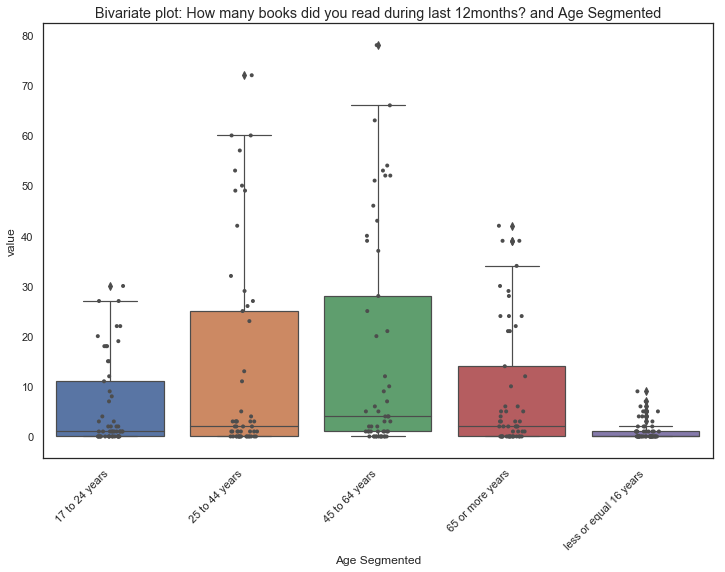
|  |  |
| --- | --- |
| Age Segmented | Number of Books (Segmented) |
| less or equal 16 years | 1 to 25 |
| 17 to 24 years | 26 to 49 |
| 25 to 44 years | 50 to 73 |
| 45 to 64 years | 74 or more |
| 65 or more years |  |

## Is Age correlated to How many books one person reads in a year?[[1]](#footnote-1)

|  |  |
| --- | --- |
| Kruskal’s test | |
| Kruskal's Statistics (unequal variances) =26.677 | **p=0.000** |

|  |  |
| --- | --- |
| Age x How many books did you read during last 12months? | |
| Pearson's correlation: | **Spearman's correlation:** |
| 0.09417 | **0.082017** |

|  |  |
| --- | --- |
| Kruskal's Null hypothesis: All sample distributions are equal.  Kruskal's test result: One or more sample distributions are not equal. (reject H0) |  |

Graphical user interface, application

Description automatically generated

|  |  |
| --- | --- |
| Chi-square’s test | |
| Kruskal's Statistics (unequal variances) =26.677 | **p=0.000** |

**Chart, bar chart

Description automatically generated**

Chi-square’s Null hypothesis: Features are independent.

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

\*Assumptions are not met in some of the number of observations individual item. Chi-square item may not be reliable. Gather more data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age Segmented \ Number of Books (Segmented) | 1 to 25 | 26 to 49 | 50 to 73 | 74 or more | All |
| 17 to 24 years | 260 | 36 | 14 | 17 | 327 |
| 25 to 44 years | 602 | 54 | 34 | 31 | 721 |
| 45 to 64 years | 660 | 76 | 59 | 60 | 855 |
| 65 or more years | 343 | 42 | 36 | 51 | 472 |
| less or equal 16 years | 50 | 8 | 4 | 4 | 66 |
| All | 163 | 147 | 216 | 1915 | 2441 |

**Conclusion:** At first look Age and **number of books read in a year** does not appear to be correlated. But, by **segmenting age** (early education, high education, early and late professional, and retirement years) a ***relationship seems to be present.***

This trend is also pointed when both Age and number of books read is segmented. Although, more data is necessary to confirm tests of independence point to a relationship.

1. P-values rounded to 3 decimals [↑](#footnote-ref-1)