Demographic characteristics' impact on reading habits

Nathália Dayse Silva do Nascimento[1]*; Douglas Augusto de Paula[2]

[1] Bachelor in Mechanical Engineering by UFPE. 101 Domville St – basement apartment – N0G 1A0 Arthur, ON, Canada
[2] Master in Controlling and Accounting by FEA/USP. Rua Costa Barros, 2050 – Bloco 5, apartamento 101 – Sítio Pinheirinho; 03210-001 São Paulo, SP, Brazil
*correspondent author: silvadon@tcd.ie

Demographic characteristics' impact on reading habits

Abstract

Considering the common sense propensity to relate reading habits with outstanding professional performance, this study proposed to explore those habits through quantitative Supervised Machine Learning methods. Therefore, General Linear Models [GLM] were compared in order to select the one that would better describe the distribution of the dependent variable from the target dataset and Count Data regression models were chosen to assess the relationship between individuals' general characteristics/status and how many books were read by them. That was done with the purpose of investigating whether or not the people's behavioural changes in literary habits would be impacted by their peculiarities. As a result, the final model was able to describe people's habits in terms of sex, education level, job status, race, income and books' formats.

Keywords: machine learning; books**;** readers; count data regression.

Introduction

Since the earliest stages of childhood, it is common for children to hear about how important the habit of Reading is, mainly at school but also from their parents. What is interesting is that even hearing the same discourse, there is no uniformity on how much those children will read once they become adults. That leads to the questions: What's the trigger for future readers? Is it really a matter of difference in personality or could other constraints/environments impact on people's reading habits? Also, will the pleasure of reading be something defined during people's childhood or is it something that can be developed throughout their lives?

After reading for their children on their initial years as a "distraction", many parents (together with teachers) encourage reading as a way to improve the young's performance at school, "to get good grades". According with Palani (2012), an outstanding educational performance demands proper reading, as this last one is connected with the total educational course. Furthermore, the common sense tends to expect that a good educational performance will drive a successful professional life in future, plus quality of life that comes with good wages. So should governments be investing more on initiatives to promote reading?

Although there is this tendency to correlate Reading and Education, it is possible that the Reading per se can drive good development. Since a study executed by Brazilian Brief Neuropsychological Assessment Battery NEUPSILIN evaluated the influence of both writing and reading habits on the neuropsychological performance of adults and their results proposed that reading and writing habits could have the potential to compensate for low education when testing cognitive abilities, such as language and attention. The testing data showed close outcomes for individuals with high frequency of reading and writing habits, but

with different levels of education (low x high) for the two mentioned cognitive tasks (Pawlowski et al., 2012).

It is not something simple to answer all the questions that were highlighted here and to prove them with actual data, as that demands not only time, but also computational power and much data that is not so easily available for public access. However, it is possible to start evaluating the possible impact of some demographic characteristics on people's current's reading habits, which is what this project aims to do.

Scales e Rhee (2001) used regression analysis to study the relationship between reading habits plus patterns with the variables gender, race, age, education level and employment and their results suggested that education and race were important predictors for these habits. Now, this study investigates if additional characteristics as income and area of residence could propose new conclusions. Also, the goal is to apply Machine Learning algorithms in a bigger database than the one used by Scales and Rhee, but also about American people. Because people's income could prove to be an important feature, the analysis in future could be used to drive new public policies to encourage reading for both children and adults.

In summary, this project aims to evaluate, through quantitative Supervised Machine Learning methods, a database on reading habits. With the objective of exploring the impact of variables such as income range, education, sex and others on determining the number of books read by individuals. Thus, it seeks to investigate whether there are behavioural changes in literary habits among people of different schooling levels, and whether the changes would be associated with their peculiarities.

Material and Methods

The chosen database was downloaded from the Pew Research Centre's website and it was a result of a telephone survey of American adults performed between March and April of 2016. The Centre consulted people about their reading habits, civic and community engagement, technology interactions, role of information in their lives and their demographic status. Later the gathered data was used to trace the profile of the population reading performance and to compare it with the scenarios of previous years. According to them, based on the percentage of readers of different educational levels interviewed in 2016, people who went to college were more likely to read books than those that did not (Pew Research, 2016).

The database was available in a CVS format, containing 1601 rows of observations and 135 variables. However, for the purpose of this project's investigation, only 21 variables were selected, focusing on individuals' demographic characteristics and their book reading

performance. Also, the original file was built in a coded form, with numbers being equivalent to different answers for each variables, which themselves were also written in codes. That created the need of using a support file to allow the dataset interpretation. That being said, Pew Centre made available a Microsoft Word document with the original questionnaire used to execute the survey, together with the codes associated to their meanings.

This leads to the next step after the data acquisition: its study and data wrangling, which was done by using the R language and environment, given its effective data manipulation capacity (R-project.org). The wrangle consists of a process that begins with the data being imported into R, followed by the steps of tidy data and transformation. That allowed the original database's "translation" into something understandable for a common reader and useful for modelling purposes (Wickham and Grolemund, 2017). In addition, outliers needed to be evaluated and missing values needed to be properly considered, this last either by replacing them by mean of other observations or completely omitting their row from the analysis (Singh et al., 2016).

Once the data was modified, this project applied Machine Learning algorithms to visualize how the amount of books read by the American population can be affected by the individuals' characteristics. Therefore, the focus was on Supervised algorithms, more specifically Generalized Linear Models (GLM), capable of predict and classify reading habits' observations based on what was learned from the patterns within existing dataset (Alloghani et al., 2019).

Variables

The questionnaire used as a base for the survey was split in two forms A and B that accounted each for 50% of the interviewed people. Nonetheless most of the questions were applied to both groups and the following data transformation considered the adjustment of variables applied to specific groups. From the original database of 135 variables, initially 21 of them were selected to compose this investigation. This decision was made considering the potential relevance of each variable to the target theme. Also, 7 of them were changed into 3 as will be detailed here.

Considering the transformed dataset with 17 selected features, below is described the exploratory analysis of each one of them, aiming to better understand the available data, before moving to the selection of the appropriate Supervised Method.

Sex (sex)

The variable sex was a categorical nominal variable that presented the answer of the interviewed people regarding whether they were male or female and it was classified in R as a factor. The Table 1 presented shows 52% of the database were men and there was zero Not Available [NA] answers.

Table 1. Frequency table for sex variable

Category	Frequency	Relative Frequency	Percentage
Male	833	0.52	52
Female	768	0.48	48
Total	1601	1	100

Source: Research's original data

Age (age)

The variable age was a numeric discrete one, representing as its name says the age of the interviewed individuals, and as so it was classified in R as an integer. Also, it is important to note that some of the people did not give an answer to this question, more specifically 2% of them. Moreover the Table 2 shows the youngest person to be part of the questionnaire had 16 years old and the oldest one had 95, plus the overall mean age between the participants was 49 years old. The minimum age was actually a limitation imposed by the survey research team as they did not interview people younger than age 16.

Table 2. Descriptive analysis of age variable

N	Mean	Standard Deviation	Min	Q1	Median	Q3	Max	Missing Values
1571	49.31	18.85	16	33	51	64	95	30

Source: Research's original data

Marital Status (marital)

The survey also questioned the participants about their current marital status at the time of the interview which was named as marital in the database. And because this was a categorical nominal variable, it was classified in R as a factor.

The Table 3 shows the majority of the individuals were married at the time of the survey or had been married some time before it (divorced and widowed).

Table 3. Frequency table for marital variable

Category	Frequency	Relative Frequency	Percentage
Married	737	0.46	46.0
Never married	397	0.25	24.8
Divorced	181	0.11	11.3
Widowed	138	0.09	8.6
Living with a partner	96	0.06	6.0
Separated	40	0.03	2.5
NA	12	0.01	0.7
Total	1601	1	100

Source: Research's original data

Parenthood Status (kid)

The kid variable (categorical nominal) was classified in R as factor and showed whether the individual was the parent or guardian of any children under age 18, considering the children were living with the interviewed person at the time of the survey. That being said, the Table 4 for this variable shows more than 75% of the people had no child at that point (at least not under 18).

Table 4. Frequency table for kid variable

Category	Frequency	Relative Frequency	Percentage
No	1205	0.753	75.3
Yes	391	0.244	24.4
NA	5	0.003	0.3
Total	1601	1	100

Source: Research's original data

Education Level (educ)

Another important demographic characteristic about the individuals was their highest level of school or the highest degree acquired by them, being a categorical nominal variable. This was named educ in the database and was classified in R as factor, having 8 different categories, as can be seen on the attached Table 5.

Table 5. Frequency table for educ variable

(continues)

Category	Frequency	Relative Frequency	Percentage
High school graduate	382	0.239	23.9
4 years college or university degree/bachelor's degree	365	0.228	22.8
Postgraduate or professional degree	247	0.154	15.4

Table 5. Frequency table for educ variable

(conclusion)

Category	Frequency	Relative Frequency	Percentage
Some college, no degree	232	0.145	14.5
2 years associate degree from college/university	158	0.099	9.9
High school incomplete	118	0.074	7.4
Less than high school	53	0.033	3.3
Some postgrad/professional schooling, no postgrad	37	0.023	2.3
NA	9	0.006	0.6
Total	1601	1	100

Source: Research's original data

Employment Status (job)

The job variable was a categorical nominal one and aimed to classify people as per their employment status, being classified in R as a factor. Additionally, 67.8% of the sample was of individuals whether employed or retired, as can be evaluated on the following Table 6.

Table 6. Frequency table for job variable

Category	Frequency	Relative Frequency	Percentage
Full time	712	0.445	44.5
Retired	373	0.233	23.3
Not for pay	218	0.136	13.6
Part-time	206	0.129	12.9
Disabled	33	0.021	2.1
Self-employed	25	0.016	1.6
Student	18	0.011	1.1
NA	16	0.010	1.0
Total	1601	1	100

Source: Research's original data

Race (race)

The survey asked people about how they would describe themselves in terms of race and the first race category listed by each person was recorded under the variable race (categorical nominal), which was classified in R as a factor. The Table 7 shows 75.5% of the interviewed people considered themselves as white.

Table 7. Frequency table for race variable

Category	Frequency	Relative Frequency	Percentage
White	1209	0.755	75.5
Black/African-american	191	0.119	11.9
Hispanic/Latino	85	0.053	5.3
Asian/Asian-american	59	0.037	3.7
NA	36	0.022	2.2
Native American/American indian/Alaska native	20	0.012	1.2
Pacific islander/Native Hawaiian	1	0.001	0.1
Total	1601	1	100

Source: Research's original data

Income Level (inc)

To understand the financial situation of the interviewed people, the survey asked them about their total annual family income from all sources (before taxes) in the previous year, that is 2015. The inc was a categorical nominal variable, classified in R as a factor and Table 8 gives details about it.

Table 8. Frequency table for inc variable

Category	Frequency	Relative Frequency	Percentage
50 to under $75,000	247	0.154	15.4
NA	196	0.122	12.2
75 to under $100,000	179	0.112	11.2
100 to under $150,000	166	0.104	10.4
$150,000 or more	158	0.099	9.9
20 to under $30,000	142	0.089	8.9
10 to under $20,000	136	0.085	8.5
30 to under $40,000	136	0.085	8.5
40 to under $50,000	122	0.076	7.6
Less than $10,000	119	0.074	7.4
Total	1601	1	100

Source: Research's original data

A transformation was performed in R in an attempt of accessing the money availability for each individual, aiming to use this new variable inc1p to evaluate the money impact on reading habits. In other words, a person with family income of $50,000 a year, for example, would be expected to have more purchasing power if that money maintained only 2 people when compared to keeping 3 or more people.

Then, to find the individual's income range, the first step was to create the variables min (imin) and max (imax) for each original income level. After that, both min and max were divided by the house variable (to be described on the next topic), creating updated min (imin1) and max (imax1) that represented how much money was available for each person on the household.

It was clear that errors were associated with this transformation, once in reality the total income would hardly be distributed equally in a family based on many different considerations. Some of those being: the difference in money need depending on the age and gender of each household member, the difference in distribution based on who on the household earned the money, the money share agreement between the members, etc. However, because these details were not provided during the interview, the simpler solution was to just consider an even share between people.

The result of the transformation showed a considerable contrast when compared to the original income variable. The last one had 10 categories (considering the NA one) while the transformation created 57 different entries. Moreover, it was important to notice that it also generated superposition between some levels and created different levels with no upper limit of income. That because by diving the “$150,000 or more” range by different amount of people, different lower limits were created for an unknown upper limit.

In summary, considering all the listed errors and uncertainties of this transformation, it was decided to use the original inc and house variables for the development of the model.

Household Members (house)

Because the income level was asked for the individuals in terms of family income, it was important to also consider on this project the amount of people living under the same household (including the interviewed person). To do this, the quantity was saved in the numeric discrete house variable, described in Table 9. Also, more or less people living under the same roof could be an important point to be taken into account when thinking about available hours to devote to reading.

The original database (1601 observations) showed a total of 14 people which answered their household had “8 or more” people. However, to allow working with this characteristic as an integer in R, the database was updated to consider all “8 or more” as 8 individuals. This strategy however added some error to the model and would impact the prediction of amount of books read on future steps of the algorithm.

Table 9. Descriptive analysis of house variable

N	Mean	Standard Deviation	Min	Q1	Median	Q3	Max	Missing Values
1583	2.65	1.5	1	2	2	4	8	18

Source: Research’s original data

Area of Residence (area)

The Pew Research Centre was also interested in classifying people based on what kind of place they lived at. This categorical nominal variable, named area, was classified in R as factor and showed only 17.5% as per Table 10 of the sample lived in a rural area, while others would concentrate in cities or close to them.

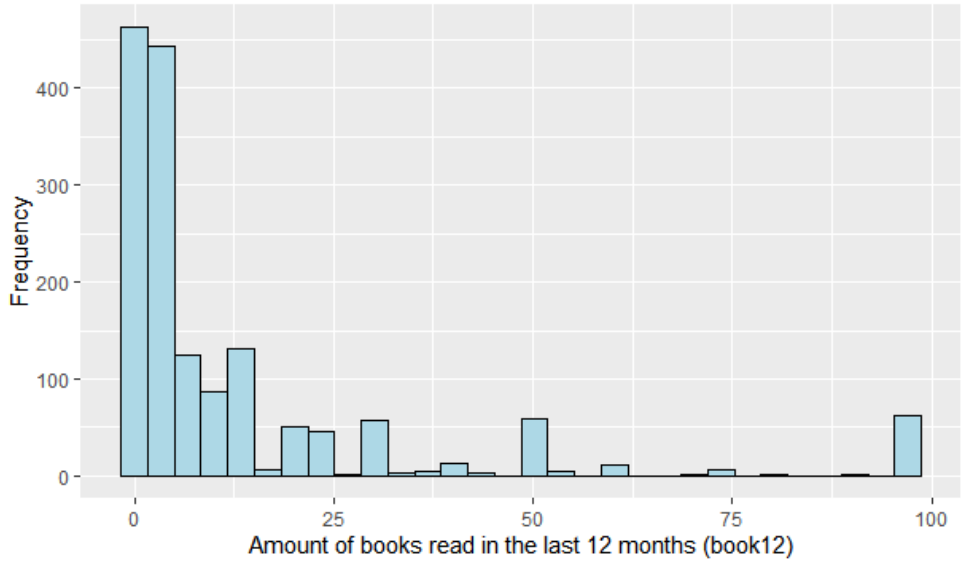
Table 10. Frequency table for area variable

Category	Frequency	Relative Frequency	Percentage
Small city or town	599	0.374	37.4
Suburb near a large city	355	0.222	22.2
Large city	353	0.220	22.0
Rural area	280	0.175	17.5
NA	14	0.009	0.9
Total	1601	1	100

Source: Research's original data

Books read in the past 12 months (book12)

The book12 variable (numeric discrete) was classified in R as an integer and it showed how many books were read in part or total by each individual in the past 12 months, showing a mean of 13 books read in a year. Picture 1 presents the histogram of its original distribution.



Picture 1. Histogram of book12 variable
Source: Research's original data

The original database (1601 observations) showed a total of 59 people which answered they had read "97 or more" books on the last 12 months. That answer format

made this variable to be classified by R as a character by default. So in order to allow its transformation to integer, the database was updated to consider all “97 or more” as 97 books, similar to what was done to the house variable. This strategy however potentially added more error to the model than the previous one, as that change was made to 0.87% of the data and this one was to 3.69% of it. Table 11 presents the descriptive analysis of book12.

Table 11. Descriptive analysis of book12 variable

N	Mean	Standard Deviation	Min	Q1	Median	Q3	Max	Missing Values
1578	13.01	21.84	0	1	4	12	97	23

Source: Research’s original data

Printed books read in the past 12 months (print12)

For the format of books read during the past year, those who said “yes” to printed books were a total of 68.7% of the interviewed people as per Table 12. Moreover, the print12 variable was categorical nominal and properly classified in R as factor.

Table 12. Frequency table for print12 variable

Category	Frequency	Relative Frequency	Percentage
Yes	1100	0.687	68.7
NA	392	0.245	24.5
No	109	0.068	6.8
Total	1601	1	100

Source: Research’s original data

Audiobooks read in the past 12 months (audio12)

Now talking about audiobooks, Table 13 shows only 15.8% of the individuals had experience with this kind of format in the past 12 months. And, same as the print12, audio12 was a categorical nominal variable and as so it was also classified in R as factor.

Table 13. Frequency table for audio12 variable

Category	Frequency	Relative Frequency	Percentage
No	957	0.598	59.8
NA	391	0.244	24.4
Yes	253	0.158	15.8
Total	1601	1	100

Source: Research’s original data

Ebooks read in the past 12 months (ebook12)

The database showed that even though ebooks were more popular than audiobooks, its percentage of readers was still not a match for the printed books, as only 30.3% said “yes” to this format. Same as the last two variables, ebook12 (categorical nominal variable) was classified in R as factor. Table 14 shows details about the frequency of each answer.

Table 14. Frequency table for ebook12 variable

Category	Frequency	Relative Frequency	Percentage
No	725	0.453	45.3
Yes	485	0.303	30.3
NA	391	0.244	24.4
Total	1601	1	100

Source: Research’s original data

Public library access frequency (lib)

The lib variable (categorical nominal) was classified in R as factor and was a combination of three questions from the reading habits survey (Pew Research, 2016):

1. “How often do you visit public libraries or bookmobiles in person?” (lib_freq)
2. “How often do you use a public library website?” (libst_frq)
3. “How often do you use a public library mobile APP?” (libapp_frq)

And for its creation the algorithm considered the higher frequency of use/visit between the three questions to become the frequency for the overall public library system access for each individual. To do so without interference of the “NA” answers, they were initially all replaced by “-1” value, and after the comparison the remaining “-1” were changed back to “NA”.

Considering those who gave an answer to the three questions above, the Table 15 showed most of the interviewed people were not frequent users of public library services.

Table 15. Frequency table for lib variable

Category	Frequency	Relative Frequency	Percentage
NA	741	0.463	46.3
Less often	541	0.338	33.8
At least once a month	179	0.112	11.2
Several times a month	86	0.054	5.4
At least once a week	54	0.034	3.4
Total	1601	1	100

Source: Research’s original data

Internet access at home (net)

The types of home internet for people who answered forms A (bbhome1f1) or B (bbhome1f2) were merged into the variable net (categorical nominal), classified in R as a factor. Also, after merging the original features, the response given by the individuals was recoded to show the existence or not of internet access at home instead of the internet type as shown in Table 16.

Table 16. Categories transformation of bbhome1f1 and bbhome1f2 into net

Internet type (bhome1f1/bbhome1f2)*	Home Internet Existence (net)
1. Dial-up	Yes
2. Higher-speed	Yes
3. Both slow-speed/dial-up and higher-speed/broadband	Yes
4. Access internet only using cell phone or tablet	Yes
5. No home internet access	No
8. Don't know	NA
9. Refused	NA

Source: Research's original data
*(Pew Research, 2016)

The descriptive analysis on Table 17 shows most of the individuals (79%) had internet access in their homes. Also, a substantial percentage didn't know how to answer or refused to do it.

Table 17. Frequency table for net variable

Category	Frequency	Relative Frequency	Percentage
Yes	1264	0.790	79.0
NA	286	0.179	17.9
No	51	0.032	3.2
Total	1601	1	100

Source: Research's original data

Local public library existence (plib1)

The impact of the closure of the local public library on the interviewed people and their families (q16a) and the impact on their community (q16b) was used to stablish if there was or not a public library on their local area. The plib1 (categorical nominal) variable was then recoded as shown in Table 18.

Table 18. Categories transformation of qb16a and qb16b into plib1

Impact of local public library closure on individuals and their family/community (qb16a/qb16b)*	Local Public Library Existence (plib1)
1. Major impact	Yes
2. Minor impact	Yes
3. No impact	Yes
4. Community does not have a public library	No
8. Don't know	NA
9. Refused	NA

Source: Research's original data
*(Pew Research, 2016)

The results in Table 19 show that almost all individuals had a public library located in their community's area.

Table 19. Frequency table for plib1 variable

Category	Frequency	Relative Frequency	Percentage
Yes	1593	0.995	99.5
NA	7	0.004	0.4
No	1	0.001	0.1
Total	1601	1	100

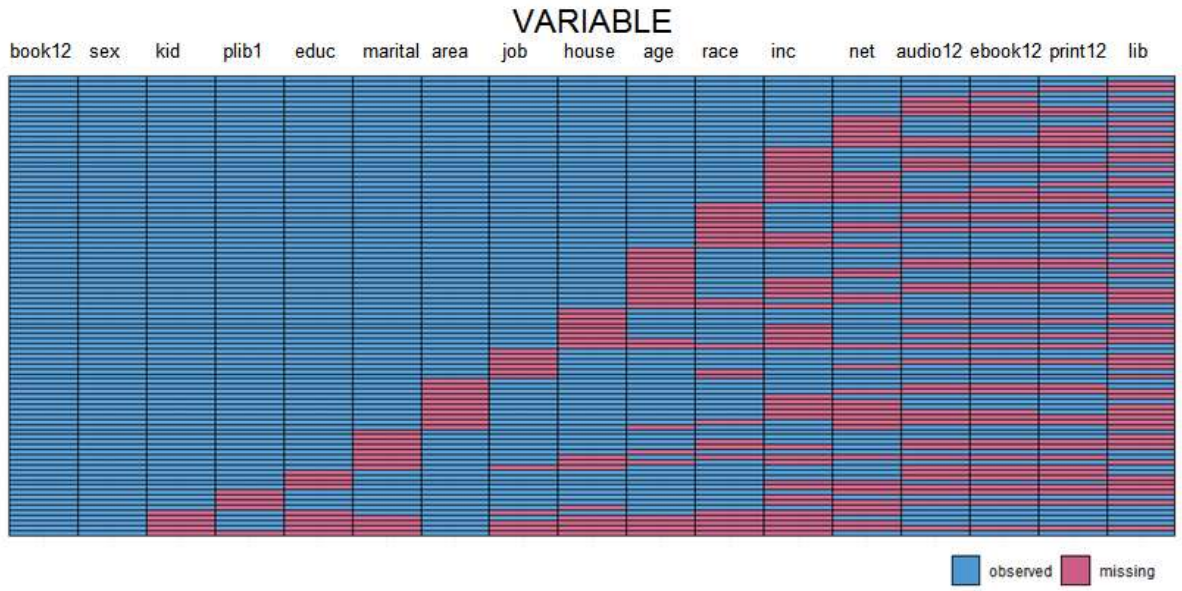
Source: Research's original data

Missing Values Treatment

When evaluating the available data, it was shown that all variables but sex presented some missing values. The reason behind it is that during the survey sometimes the interviewed people either would not know how to answer specific questions or they refused to answer them. Then, it led to the need of treating the database for missing data.

Because book12 was chosen as the dependant variable for this project and its percentage of missing values was only 1.44%, it was decided to use drop_na() to remove from the dataset all lines which did not present a value for this variable. That been done, the new dataset had 1578 observations.

Then, before moving forward with any additional data missing treatment, the package ggmice was used to visually evaluate the dataset for any kind of patterns on its missing values. Which given how dispersed the pink cells (missing values) are on the Picture 2, can be classified as a General Missing Data Pattern (Enders, 2010). Also, both Picture 2 and Table 20 show a high percentage (46.39%) of NA occurrences for the lib variable, so, to avoid introducing bias to the data by trying to impute values, it was decided to remove this column from the investigation.



Picture 2. Plot of the missing data pattern of books survey dataset
Source: Research's original data

Table 20. Percentage of Missing Values after dropping the NA lines of book12

Variable	Missing Values (%)
Lib	46.39%
print12	23.38%
audio12	23.32%
ebook12	23.32%
Net	17.74%
Inc	12.17%
Race	2.15%
Age	1.90%
House	1.14%
Job	1.01%
Area	0.89%
Marital	0.76%
Educ	0.57%
plib1	0.44%
Kid	0.32%
Sex	0.00%
book12	0.00%

Source: Research's original data

Here, due to its convenience, for the remaining missing values it was decided to follow a single imputation procedure, that means to fill the missing data with one value to transform the dataset in a complete one (Jamshidian and Mata, 2007). However, it is important to highlight that its application has both pros and cons: e.g. it allows using data that would not be available when using the deletion techniques, but it underestimates sampling error (Enders, 2010). Here for numeric variables the missing values were replaced by the

features' mean (rounded value as they were all discrete) and for categorical ones they were replaced by the columns' mode.

Potential Outliers Identification

For the three numeric variables (book12, age and house) the dataset was evaluated for the existence of univariate outliers. Boxplots were created to visually guide the identification of any suspicious values and the Interquartile range (IQR) described by equation (1) was used to classify the points as moderate or extreme outliers (Fávero and Belfiore, 2017).

$$IQR=Q_3-Q_1 \quad (1)$$

where, Q_1: is the first quartile; and Q_3: is the third quartile.

Considering the value x of a variable, it was classified as moderate outlier when equation (2) or (3) was true.

$$x<Q_1-1.5 \times IQR \quad (2)$$

$$x>Q_3+1.5 \times IQR \quad (3)$$

And it was classified as extreme outlier when equation (4) or (5) was true.

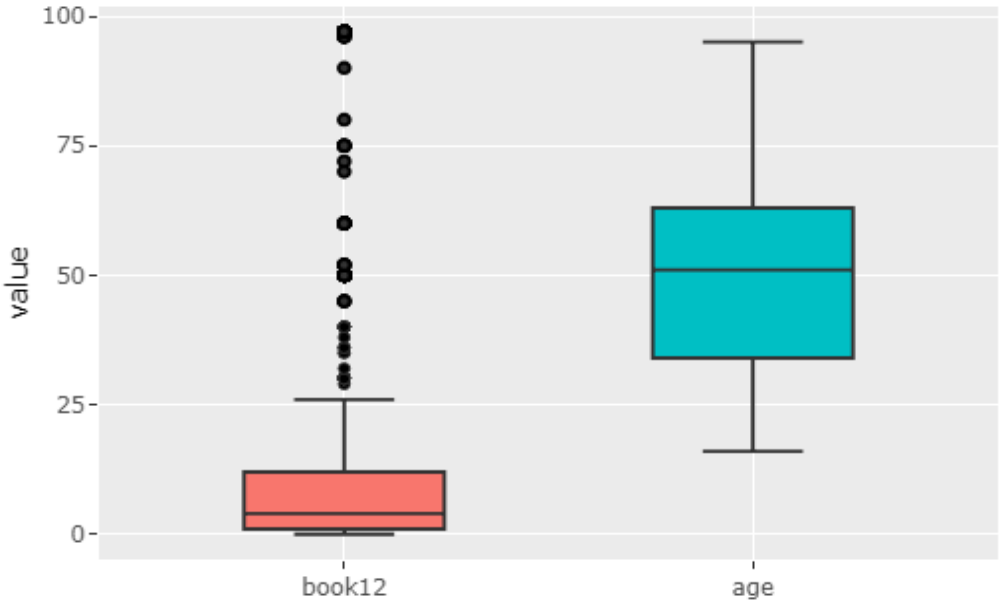
$$x<Q_1-3 \times IQR \quad (4)$$

$$x>Q_3+3 \times IQR \quad (5)$$

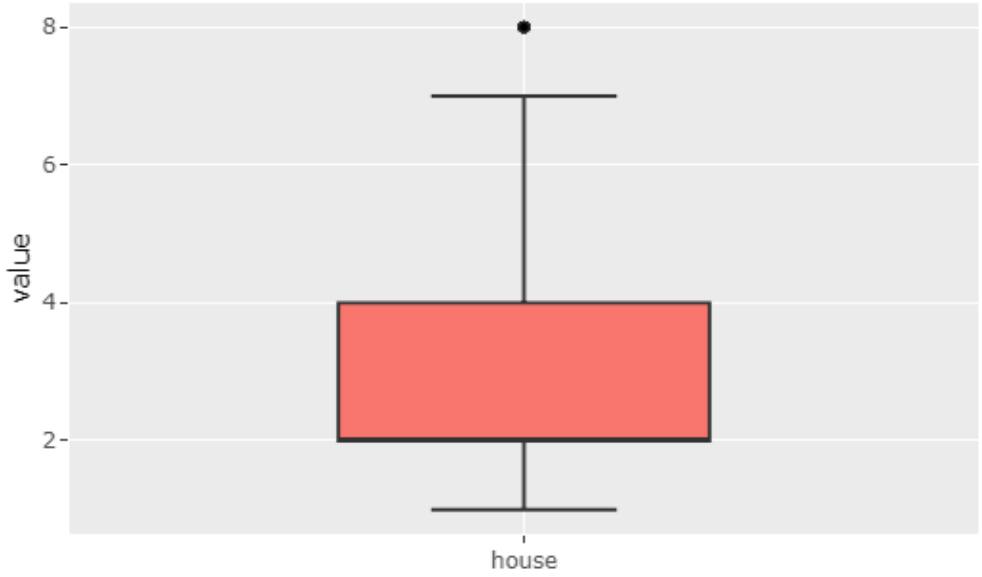
Picture 3 shows the boxplot for the variables book12 and age, while Picture 4 shows the one for the house variable. Additionally Table 21 shows the results for the IQR and classification ranges for each variable. The analysis of both pictures and table shows age had no outliers, home had 14 occurrences identified as moderate and book12 had 23 moderate occurrences and 146 occurrences (9.3% of the database) of potential extreme outliers.

These potential outliers could be mainly a result of misunderstanding question's wording, its response or even typo issues. However, a quick look on reading forums online can show that reading 97 books in a year is not something unreasonable as some people may think. Considering this, plus the fact that the potential extreme points represent a

significant percentage of the data and this project's goal to understand who reads more often, were all reasons to initially do not remove the 146 occurrences from the dataset.



Picture 3. Boxplot of book12 and age
Source: Research's original data



Picture 4. Boxplot of house
Source: Research's original data

Table 21. IQR Results

Variable	IQR	Outlier Classification	Range
book12	12	Moderate	x<-17 or x>31
		Extreme	x<-35 or x>49
age	29	Moderate	x<-9.5 or x>106.5
		Extreme	x<-53 or x>150
house	2	Moderate	x<-1 or x>7
		Extreme	x<-4 or x>10

Source: Research's original data

Modelling

The model was developed using the R language, with focus on Supervised Machine Learning Methods and having the annual amount of ridden books (book12) as the dependent variable. Considering most of the selected variables were qualitative, the function dummy_columns from package fastDummies was used to turn each categorical variable into (n-1) dummies, being n the number of categories of each feature. This created a new database with 44 variables and the reference categories were the most frequent ones as per Table 22. Because books12 was a discrete feature with no negative occurrences and with exposition of 12 months, Count Data Regression analysis was chosen for the model.

Table 22. Original variables and its reference category for dummy transformation

Original Variable	Reference Category
Sex	Male
Marital	Married
Kid	No
Educ	High school graduate
Job	Full time
Race	White
Inc	$50 to under $75,000
Area	Small city or town
print12	Yes
audio12	No
ebook12	No
Net	Yes
plib1	Yes

Source: Research's original data

The Picture 1 shown previously presented a histogram with a long tail, which together with the high difference between the book12's mean (13.01) and its variance (476.79) were strong signs of the overdispersion phenomenon. To confirm, the Cameron and Trivedi (1990) overdispersion test was performed using the overdisp() algorithm and it returned a p-value of $2.2e^{-16}$ (Souza et al., 2022). That suggested the Gamma Poisson (Negative Binomial) probability distribution shown in equation (6) would be a better fit for the book12 behaviour than the Poisson distribution (Fávero and Belfiore, 2017).

.

$$p(Y_i=m)=\frac{\delta^{\theta}.m_i^{\theta-1}.e^{-m_i.\delta}}{(\theta-1)!}\ ,\ i=1, 2, 3\ldots \quad (6)$$

where Y is the dependent variable, i represents each case of the sample, θ is the shape parameter and δ is the rate parameter.

To establish the parameters α and β's for the Gamma Poisson model presented by equation (7), the logarithm of the likelihood (LL) function (8) was optimized using function glm.nb() from MASS package followed by the Stepwise method (function step()) to find the maximum LL (Fávero et al., 2021).

$$\ln(\hat{Y}_i) = \ln\left(\lambda_{bneg_i}\right) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} \ , \ i=1,2,3\ldots \tag{7}$$

$$LL = \sum_{i=1}^{n}\left[Y_i.\ln\left(\frac{\emptyset.\lambda_{bneg_i}}{1+\emptyset.\lambda_{bneg_i}}\right) - \frac{\ln\left(1+\emptyset.\lambda_{bneg_i}\right)}{\emptyset} + \ln\Gamma\left(Y_i+\emptyset^{-1}\right) - \ln\Gamma\left(Y_i+1\right) - \ln\Gamma\left(\emptyset^{-1}\right)\right] = max \tag{8}$$

where, X_1 to X_k are explanatory variables and Ø is the inverse of the shape parameter θ.

Additionally to the Gamma Poisson model, the reasonable percentage of zero occurrences of book12 (23.1% of the database) indicated it would be valid to investigate a Zero Inflated Negative Binomial [ZINB] model for its behaviour. So, given the fact that currently there is no Stepwise function in R for ZINB regressions, the variables to be used on the model had to be pre-selected (Fávero et al., 2021). The approach chosen followed below steps:

1. Created a new variable based on book12, where bookbi assumed "yes" value if at least one book was read in the last 12 months and assumed "no" otherwise;
2. Estimated a Binary Logistic model to explain bookbi behaviour with glm(family="binomial") function;
3. Applied the step() function to the Binary model;
4. Estimated the ZINB model using function zeroinfl() with only the final variables selected on the stepwise from the Gamma Poisson model and the ones selected on the stepwise from the Binary Logistic model (step 3).

   When doing that, it was assumed the equation (9) distribution for book12.

$$\begin{cases} p(Y_i=0) = p_{logit_i} + \left(1-p_{logit_i}\right).\left(\dfrac{1}{1+\theta^{-1}.\lambda_{bneg_i}}\right)^{\theta} \\ p(Y_i=m) = \left(1-p_{logit_i}\right).\left[\dfrac{\delta^{\theta}.m_i^{\theta-1}.e^{-m_i.\delta}}{(\theta-1)!}\right] \ , \ m=1, 2, 3\ldots \end{cases} \tag{9}$$

$$p_{logit_i} = \frac{1}{1+e^{-(\gamma+\delta_1.W_{1i}+\delta_2.W_{2i}+\ldots+\delta_q.W_{qi})}} \tag{10}$$

$$\lambda_{bneg_i} = e^{(\alpha+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_k X_{ki})} \quad (11)$$

where the parameters α, β's, γ and W's were chosen by optimizing the logarithm of the likelihood function (12) and were used to estimate the ZINB model in equation (13).

$$LL = \sum_{Y_i=0} \ln\left[p_{logit_i} + \left(1-p_{logit_i}\right).\left(\frac{1}{1+\emptyset.\lambda_{bneg_i}}\right)^{\frac{1}{\emptyset}}\right] + \sum_{Y_i>0}\left[\begin{array}{c}\ln\left(1-p_{logit_i}\right)+Y_i.\ln\left(\frac{\emptyset.\lambda_{bneg_i}}{1+\emptyset.\lambda_{bneg_i}}\right)- \\ \frac{\ln\left(1+\emptyset.\lambda_{bneg_i}\right)}{\emptyset}+\ln\Gamma\left(Y_i+\emptyset^{-1}\right)- \\ \ln\Gamma(Y_i+1)-\ln\Gamma(\emptyset^{-1})\end{array}\right] = max \quad (12)$$

$$\widehat{Y}_i = \lambda_{ZINB} = \left\{1 - \frac{1}{1+e^{-(\gamma+\delta_1.W_{1i}+\delta_2.W_{2i}+\ldots+\delta_q.W_{qi})}}\right\}.\left\{e^{(\alpha+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_k X_{ki})}\right\} \quad , \; i=1,2,3\ldots \quad (13)$$

Results and Discussion

The stepwise method was used to stablish the best Gamma Poisson model based on the evaluation of the statistical significance of the available independent (explanatory) variables at the significance level of 5%, in a way to find the equation with smallest AIC. From the original 43 features, only 13 (all dummies) were selected for the model presented by equation (14).

$$\widehat{book12} = e^{\left(\begin{array}{c}1.72+0.55*fem+0.19*ed1-0.95*ed2+0.31*ed3+0.25*jb1+0.28*jb2-0.71*jb3 \\ -0.28*rc1-0.74*rc2+0.27*inc1-0.40*prt+0.61*aud+0.69*ebk\end{array}\right)} \quad (14)$$

It is important to note that the intercept 1.723 carries the behaviour of the reference categories chosen during the dummy transformation, listed previously on Table 22. The variables' selection for this model is detailed on Table 23, plus the fact that these features together built the best model doesn't mean that all the ones removed had no capacity to explain some parcel of the book12's variance.

Table 23. Estimated count data regression models for the number of books read in a year

(continues)

	Count data Regression			
	Gamma Poisson		ZINB	
Explanatory variables	Coefficient	p-value	Coefficient	p-value
α	1.723	$2.00e^{-16}$ ***	2.126	$2.00e^{-16}$ ***
X's (Count data model)				
sex				
Female (fem)	0.549	$2.06e^{-13}$ ***	0.447	$1.22e^{-10}$ ***
educ				

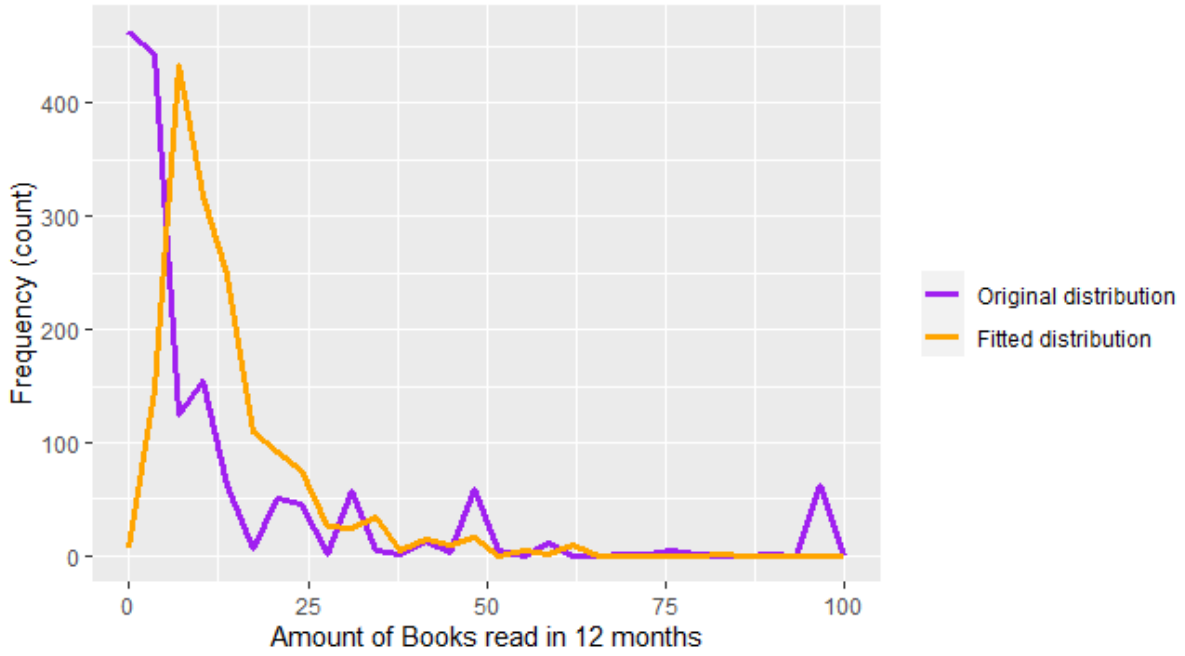
Table 23. Estimated count data regression models for the number of books read in a year (conclusion)

	Count data Regression			
	Gamma Poisson		ZINB	
Explanatory variables	Coefficient	p-value	Coefficient	p-value
4 years college / university or bachelor degree (ed1)	0.191	0.041 **	0.092	0.278
Less than high school (ed2)	-0.948	$2.15e^{-05}$ ***	-0.610	0.016 **
Post-graduation or professional degree (ed3)	0.306	0.006 ***	0.141	0.142
job				
Part time job (jb1)	0.253	0.026 **	0.104	0.308
Retired (jb2)	0.280	0.002 ***	0.301	0.000 ***
Self-employed (jb3)	-0.711	0.023 **	-0.453	0.143
race				
Black or African-American (rc1)	-0.278	0.017 **	-0.276	0.011 **
Hispanic/Latino (rc2)	-0.740	$2.58e^{-05}$ ***	-0.539	0.003 ***
inc				
$150,000 or more (inc1)	0.267	0.037 **	0.219	0.060 *
print12				
Person did not read printed books in the last year (prt)	-0.401	0.008 ***	-0.525	$2.30e^{-05}$ ***
audio12				
Person listened to audio books in the last year (aud)	0.611	$5.28e^{-09}$ ***	0.413	$1.70e^{-06}$ ***
ebook12				
Person read ebooks in the last year (ebk)	0.687	$4.06e^{-15}$ ***	0.485	$6.76e^{-11}$ ***
γ			-0.136	0.418
W's (Zero-inflation model)				
sex				
Female (fem)			-0.739	0.000 ***
educ				
4 years college / university or bachelor degree (ed1)			-0.888	0.001 ***
Less than high school (ed2)			0.705	0.096 *
Post-graduation or professional degree (ed3)			-2.069	0.002 ***
2 year associate degree from college/university (ed4)			-0.679	0.051 *
Some college / no degree (ed5)			-0.741	0.013 **
job				
Part time job (jb1)			-0.901	0.017 **
Student (jb4)			-19.563	0.997
race				
Person is Hispanic/Latino (rc2)			0.729	0.055*
inc				
$10,000 to under $20,000 (inc2)			0.628	0.032 **
$100,000 to under $150,000 (inc3)			-0.705	0.110
print12				
Person did not read printed books in the last year (prt)			-18.295	0.994
audio12				
Person listened to audio books in the last year (aud)			-18.516	0.992
ebook12				
Person read ebooks in the last year (ebk)			-18.979	0.990
Number of variables		13		18
Log-Likelihood (LL)		-5209		-5103
LL Ratio Test (vs Gamma) – p-value				$2.20e^{-16}$ ***
Vuong Test (vs Gamma) raw – p-value				$8.80e^{-15}$ ***

Source: Research's original results

*** Significant at 1%; ** Significant at 5%; * Significant at 10%

The created Gamma Poisson model was plotted together with the book12's frequency distribution from the dataset and the result is shown in Picture 5. It illustrated relatively a better adherence to the original curve on the range from 38 to 94 books and unfortunately a poor performance in predicting the most frequent readers. Who is important to notice would be the suggested outliers based on the IQR method. Also, the orange curve seemed to present a slight lag to the right when compared to the original one for those who read less than 20 books.



Picture 5. Comparison between original frequency distribution of variable book12 and fitted distribution from Poisson Gamma model
Source: Research's original results

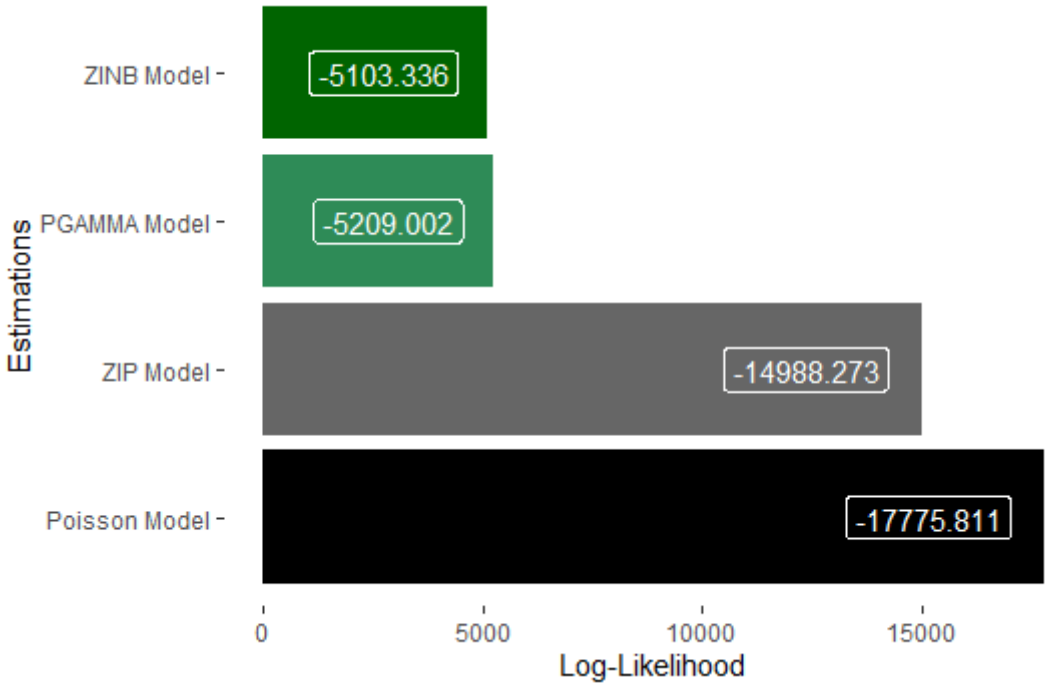
The ZINB model estimation considered 18 variables, some of them being used to describe both the count of occurrences, acting as X's features, and the occurrence of structural zeros, acting as W's features. The list of variables and their respective coefficients were presented on Table 23 and the resulting model is described in equation (15).

$$\widehat{book12} = \left\{1 - \frac{1}{1+e^{\begin{array}{c}-(-0.14-0.74.fem-0.89.ed1+0.71.ed2\\-2.07.ed3-0.68.ed4+0.74.ed5-0.90.jb1\\-19.56.jb4+0.73.rc2+0.63.inc2-0.71.inc3\\-18.30.prt-18.52.aud-18.98.ebk)\end{array}}}\right\} . \left\{e^{\left(\begin{array}{c}2.13+0.45.fem+0.09.ed1-0.61.ed2\\+0.14.ed3+0.10.jb1+0.30.jb2\\-0.45.jb3-0.28rc1-0.54rc2+0.22.inc1\\-0.53.prt+0.41.aud+0.49.ebk\end{array}\right)}\right\} \quad (15)$$

Same as in equation (14), the intercepts γ and α were here carrying the behaviour of the reference categories from Table 22. However, for this second model it is important to

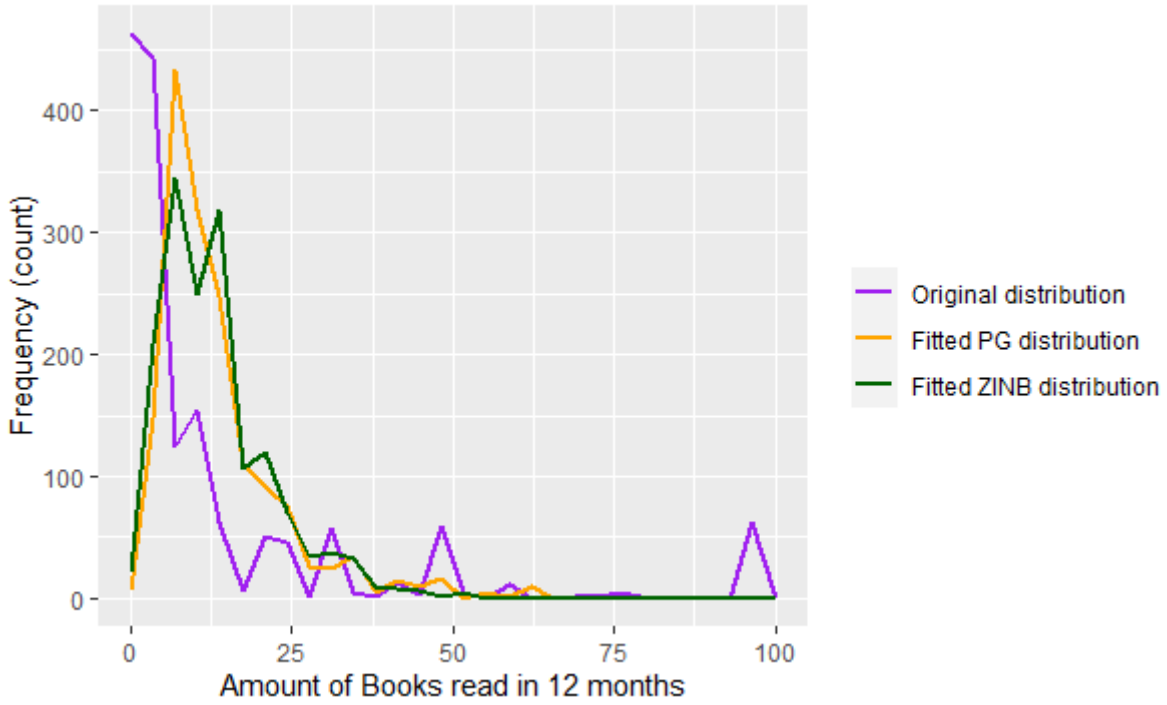
note that the result returned by R presented some parameters of predictor variables that were not statistically different from zero at the level of significance of 5%. Therefore, this could indicate that the method chosen on the previous section to select the variables for the ZINB regression model (on the absence of a stepwise function) was not efficient enough.

Nevertheless, the Vuong Test (Vuong, 1989) was evaluated with the vuong() function to compare the Poisson Gamma model and the ZINB model, which confirmed the zero-inflation, given the calculated raw p-value of $8.80e^{-15}$. The log-likelihoods of the four Count Data regression models were compared on Picture 6, which shows the ZINB as the best model, presenting a slightly higher LL (-5103.336). Also, the p-value of $2.2e^{-16}$ achieved with the log-likelihood rate (LR) test between the Poisson Gamma and ZINB models reinforced the significant difference between the LL of these two models, for a confidence level of 95%.



Picture 6. Comparison between Log-likelihood of Poisson, Zero Inflated Poisson (ZIP), Poisson Gamma and ZINB models
Source: Research's original results, printed using Program in R from Fávero et al. (2021)

The ZINB model was added to the plot with the book12's original frequency distribution and the Poisson Gamma one, in Picture 7. Similarly to the first model, the new one also presented a poor prediction of the people who read most books in a year. Additionally, few ranges, like from 0 to 12 books, had results slightly closer to the original curve than the Poisson ones had.



Picture 7. Comparison between original frequency distribution of variable book12 and fitted distribution from Poisson Gamma and ZINB models
Source: Research's original results

The ZINB's parameters statistically different from zero strongly suggest some conclusions about the behaviour of readers. On the impact of the explanatory variables on the occurrence of structural zeros, when the person is a woman, then exp(-0.74)=0.48, which implies a decrease of 52%, on average, in the chances, ceteris paribus. While in terms of education level, if the person has a 4 years degree, then exp(-0.89)=0.41, indicating a decrease of 59%, on average, in the chances of occurrence of structural zeros, ceteris paribus. However, if the person has a post-graduation or professional degree, then exp(-2.07)=0.13, suggesting a decrease of 87%, on average, in the chances of occurrence of structural zeros, ceteris paribus. Nonetheless, if the person has a 2 years degree, then exp(-0.68)=0.51, which means a decrease of 49%, on average, in the chances of the occurrence of structural zeros, ceteris paribus. Lastly, when the person has some college, then exp(0.74)=2.10, leading to an increase of 110%, on average, in the chances of occurrence of structural zeros, ceteris paribus.

Regarding other characteristics, the chances of occurrence of structural zeros would decrease 59%, on average, for people with part-time jobs, as exp(-0.90)=0.41. Additionally, when the person has a low income level, between $10,000 to $20,000 annual, then exp(0.63)=1.88, suggesting an increase of 88%, on average, in the chances of the event of structural zeros, ceteris paribus.

Now, considering the impact of the independent variables in the chance of occurrence of books read in a year, if the person is a woman, then exp(0.45)=1.57, representing an increase of 57%, on average, in the chances. However, for people with education level less than high school, then exp(-0.61)=0.54, implicating a decrease of 46%, on average, in the chances of occurrence of books read. Still, when people are retired, then exp(0.30)=1.35, leading to an increase of 35%, on average, in the chances of occurrence of books read.

Additionally, in terms of race identification, if the person self considers herself as Black or African-American, then exp(-0.28)=0.76, which implies a decrease of 24%, on average, in the chances of occurrence of books read, while for Hispanics or Latinos, exp(-0.54)=0.58 and the decrease in the chances would be of 42%, on average.

Moreover, if the person has not read printed books in the last year, then exp(-0.53)=0.59, which means that in that year her chances of reading books would decrease 41%, on average. However, if she/he has listened to audio books in the last 12 months, then exp(0.41)=1.51, leading to an increase of 51%, on average, in the chances of occurrence of books read. Furthermore, when the person has read ebooks in the last year, then exp(0.49)=1.63, representing an increase of 63%, on average, in the chances of occurrence of books read in a year.

Conclusion

This study assesses the relationship between individuals' general characteristics/status and how many books are read by them. That is done with the purpose of investigating whether or not the people's behavioural changes in literary habits would be related with their peculiarities. To do so, GLM models are evaluated in order to find which one would better describe the distribution of the dependent variable from the chosen dataset.

Because the amount of books read in a year is a discrete variable, the Count Data regression models are considered and the Cameron and Trivedi (1990) test gives support to prove the presence of overdispersion in the data, followed by the Voung (1989) test that confirms the zero inflation in the variable. Therefore, the ZINB model is the most appropriate for the prediction of books read in a year.

On one hand, the final model is able to describe the reading habits of people in terms of their sex, of their education level, of some aspects of their job status, of some race classifications, of few levels of income and the format of the read books (printed, electronic or audiobooks). On the other hand, in the presence of these variables, characteristics as age, marital status, whether the person has children or not, how many people lives in the same house, their area of residence, whether they have internet access or not and whether

their community has a public library or not, were factors that did not present statistical significance to compose the model. Still, it is important to reinforce that this is not a proof that the not considered variables are not at all able to individually influence the reading habits of individuals.

Furthermore, comparing the variables selected for the ZINB model, the study shows that people's income presents only influence on whether they would read books or not, but no impact on how often they would do it, as its statistical significance is only proven for the occurrence of structural zeros. In contrast, people's race and book formats present statistical impact only on how often people read, but do not dictate whether they would read or not. Additionally, points as sex, education and job status are able to drive both the occurrence and the count of books read in a year.

Lastly, the current non-existence of stepwise procedure in R for the optimization of Zero Inflated (ZI) models makes it complicated the proper selection of variables that are in truth statistically significant in the prediction of this group of models. That being said, there is still room for improvement on the ZINB regression described on this study, which can drive future projects.

Acknowledgment

This work would not have been possible without the support of my parents, Dilson and Cristiane, and my loving husband, Thiago, who encouraged me throughout not only this project, but also throughout the entire Data Science & Analytics MBA. Thank you!

References

Alloghani, M.; Al-Jumeily, D.; Mustafina, J.; Hussain, A.; Aljaaf, A.J. 2019. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science, 3 – 21. In: Berry, M.; Mohamed, A.; Yap, B. 2020. Supervised and Unsupervised Learning for Data Science: Unsupervised and Semi-Supervised Learning. 1ed. Springer, Conway, Arkansas, USA.

Cameron, A. C.; Trivedi, P. K. 1990. Regression-based tests for overdispersion in the Poisson model. Journal of Econometrics, v. 46, n. 3, p. 347-364.

Enders, C. K. 2010. Applied missing data analysis. 1ed. The Guilford Press, New York.

Fávero, L. P., Belfiore, P. 2017. Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. 1ed. Elsevier Editora Ltda., Brasil.

Fávero, L. P.; Souza, R. D. F.; Belfiore, P.; Corrêa, H. L.; Haddad, M. F. 2021. Count data regression analysis: concepts, overdispersion detection, zero-inflation identification, and applications with R. Practical Assessment, Research, and Evaluation, 26(1), 13.

Jamshidian, M.; Mata, M.. 2007. Advances in analysis of mean and covariance structure when data are incomplete, 21-44. In: Lee, S. 2007. Handbook of latent variable and related models. North-Holland. 1ed. Elsevier B.V, North Holland.

Josse, J.; Prost, N.; Scornet, E.; Varoquaux, G.. 2019. On the consistency of supervised learning with missing values. arXiv preprint arXiv:1902.06931.

Palani, Kumar K. 2012. Promoting reading habits and creating literate society. Researchers world 3-2: 90.

Pawlowski, J.; Remor, E.; de Mattos Pimenta Parente, M.A. et al. 2012. The influence of reading and writing habits associated with education on the neuropsychological performance of Brazilian adults. Read Writ 25: 2275–2289.

Pew Research Center [PRC]. 2016. Book Reading 2016. Available at: <https://www.pewresearch.org/internet/2016/09/01/book-reading-2016/>. Accessed at: 17 mar. 2022.

Scales, A. M.; Rhee, O. 2001. Adult reading habits and patterns. Reading Psychology 22-3: 175-203.

Singh, A.; Thakur, N.; Sharma, A. 2016. A review of supervised machine learning algorithms. 3rd International Conference on Computing for Sustainable Global Development (INDIACom):1310-1315.

Souza, R. D. F.; Fávero, L. P.; Belfiore, P.; Corrêa, H. L. 2022. overdisp: an R package for direct detection of overdispersion in count data multiple regression analysis. International Journal of Business Intelligence and Data Mining, 20(3), 327-344.

The R Foundation [TRF]. What is R? Available at: <https://www.r-project.org/about.html>. Accessed at: 3 apr. 2022.

Vuong, Q. H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica: Journal of the Econometric Society, 307-333.

Wickham, H.; Grolemund, G. 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. 1ed. O'Reilly Media, Sebastopol, CA.