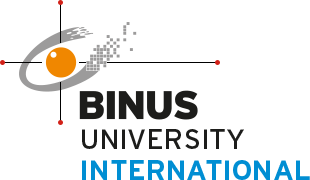
Odd Semester (2021)



# **BINUS UNIVERSITY**

# **BINUS INTERNATIONAL**

| Final Project Cover Letter(Group Work) |
| --- |

**Student Information: Surname: Given Name: Student ID Number:**

1. Clarin Maria 2501990331

2. Wijaya Rachel Anastasia 2502009646

3. Staniswinata Stephanie 2501997836

**Course Code :** STAT6171001 **Course Name :** Basic Statistics

**Class :** L3AC **Lecturer :** Raymond Bahana, ST., M.Sc

**Type of Assignments:** Term Final Project

**Submission Pattern**

**Due Date :** 9 January 2022 **Submission Date :** 9 January 2022

The assignment should meet the below requirements.

1. Assignment (hard copy) is required to be submitted on clean paper, and (soft copy) as per lecturer’s instructions.
2. Soft copy assignment also requires the signed (hardcopy) submission of this form, which automatically validates the softcopy submission.
3. The above information is complete and legible.
4. Compiled pages are firmly stapled.
5. Assignment has been copied (soft copy and hard copy) for each student ahead of the submission.

**Plagiarism/Cheating**

BiNus International seriously regards all forms of plagiarism, cheating, and collusion as academic offenses which may result in severe penalties, including loss/drop of marks, course/class discontinuity, and other possible penalties executed by the university. Please refer to the related course syllabus for further information.

**Declaration of Originality**

By signing this assignment, I understand, accept, and consent to BiNus International terms and policy on plagiarism. Herewith I declare that the work contained in this assignment is my own work and has not been submitted for the use of assessment in another course or class, except where this has been notified and accepted in advance.

Signature of Student:

1. Maria Clarin
2. Rachel Anastasia Wijaya
3. Stephanie Staniswinata

# Background

When it comes to reading books, people are always in search of books that are popular or books that are deemed ‘good’ books. This is natural because every reader wants the best reading experience when reading a book. With the growth of technology and the reading community, there are now websites in which people can rate and also view ratings of books. Books with high ratings are generally satisfactory while books with low ratings are deemed not worth a read. The current biggest online platform for readers and book recommendations are Goodreads [1]. As the largest growing book reading and recommendation platform across the Internet, Goodreads provides various information in regards to each book–the book cover, synopsis, current book rating, awards (if any), genre, author, author rating, etc. This is the information that a reader will put into consideration when determining their next read [2].

Readers have the tendency to choose their next book based on the author. Saarinen and Vakkari (2013) explained the reader's behavior of searching or browsing well-known book authors when looking for the next book to read in libraries [3]. In book communities, avid readers often have a go-to author that they like–may it be because of their writing style or the genre of books they write, readers have their own preferences on book authors. But generally, the current public perception is that the more popular an author is, the better their books are. This is because in order for an author to be popular, their books have to be popular first, meaning that their books have to be good in order to reach that popularity. This allows popular authors to be able to publish more books, as they are more likely to sell and gain popularity faster than the books written by newcomer authors. This creates a general hypothesis of the more distinct works an author has written, the better their ratings will be due to their popularity. However, as it is just a perception, we wanted to dig deeper and find the statistical proof of this hypothesis to prove whether or not it is true. The goal of our research is to prove whether the amount of distinct works an author has written has an effect on the author’s average rating or if they are independent of each other. In order to find the answer to our research, we used several statistical equations to determine the final hypothesis. These statistical equations include the Chi-squared test of independence, Pearson’s correlation coefficient, and ANOVA.

# Hypothesis

For this research, we have two main hypotheses–the null hypothesis and the alternative hypothesis. The hypothesis is as below:

* **Null hypothesis**: An author’s average rating and their number of distinct works are independent of each other.
* **Alternative hypothesis**: An author’s number of distinct works affects the author’s average rating.

The null hypothesis states that an author’s average rating will not be affected by the number of distinct works the author has written. Meaning that if the author were to write more books, their personal average rating will not be affected positively or negatively as they are independent of each other.

The alternative hypothesis states the correlation between the number of authors’ distinct works and their personal average rating. The number of an author’s distinct works can positively or negatively affect the author’s personal rating, meaning that they are dependent on each other to an extent.

# Data

The data set that we used for this final project is attached in the google sheet link before references. The data consists of the author’s details. This data set was organically scraped and wrangled from the Goodreads website data was collected from Goodreads website. The data set consists of the author’s name, the author’s average rating, and distinct works (columns). The total number of data points (referring to the number of rows) is 1238. The plot of the data set is as visualized in the box and whisker plot and normal distribution graphs below:

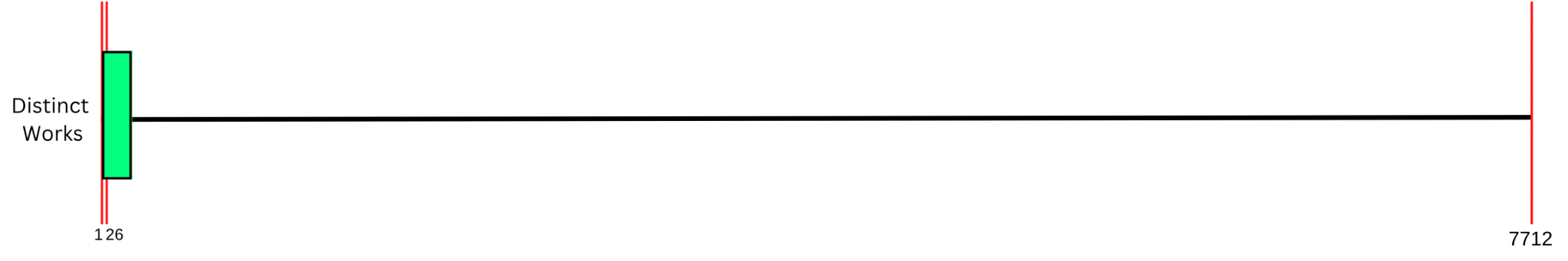
1. **Box and Whisker Plot**

**Table 3.1.** Distinct Work Box and Whisker Values

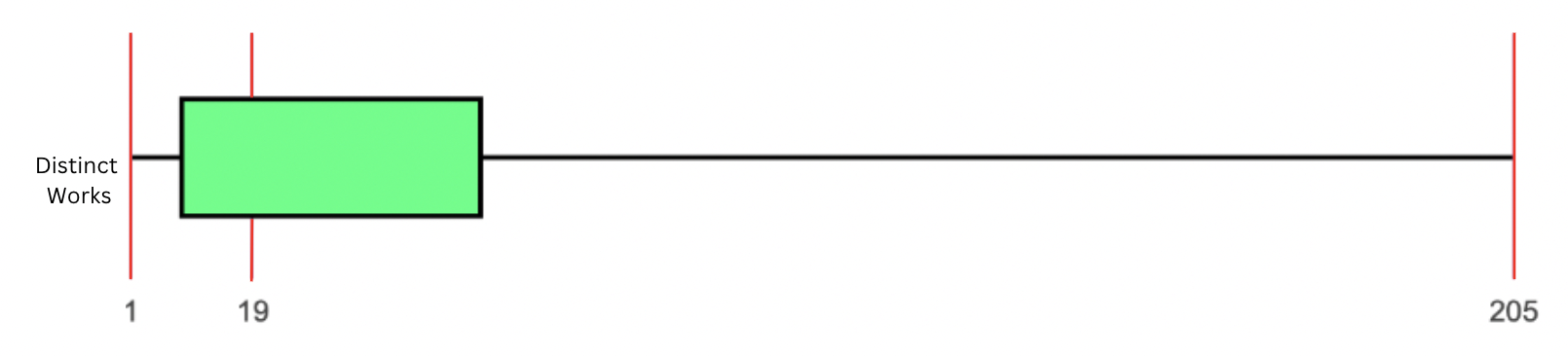
| Sample Size (n) | 1238 |
| --- | --- |
| Minimum | 1 |
| Maximum | 7712 |
| Median | 26 |
| Lower Quartile (Q1) | 9 |
| Upper Quartile (Q3) | 88 |
| Interquartile Range (IQR) | = Q3 - Q1  = 88 - 9  = 79 |
| Mean | 183.44830371567 |
| Outlier Values | Lower Limit = Q1 - 1.5×IQR = -109.5  Upper Limit = Q3 + 1.5×IQR = 206.5  Values above the upper limit or below the lower limit are considered outliers.  207, 210, 210, 211, 212, 213, 213, 214, 216, 220, 220, 221, 223, 223, 223, 225, 225, 227, 231, 233, 235, 235, 237, 240, 248, 256, 275, 278, 281, 286, 295, 308, 323, 328, 337, 340, 342, 345, 347, 352, 360, 364, 372, 372, 385, 386, 391, 403, 415, 417, 418, 425, 429, 439, 439, 440, 442, 450, 456, 461, 463, 464, 465, 470, 489, 505, 506, 512, 524, 533, 535, 543, 551, 554, 564, 573, 573, 578, 605, 634, 634, 637, 647, 648, 650, 651, 652, 668, 671, 674, 684, 693, 695, 714, 717, 729, 735, 736, 738, 746, 753, 753, 754, 796, 806, 809, 811, 838, 840, 841, 852, 861, 869, 875, 888, 913, 914, 919, 944, 954, 959, 968, 975, 1002, 1011, 1033, 1046, 1056, 1101, 1104, 1112, 1118, 1121, 1134, 1144, 1156, 1187, 1189, 1203, 1229, 1246, 1273, 1299, 1324, 1378, 1398, 1407, 1456, 1471, 1481, 1623, 1624, 1631, 1638, 1691, 1788, 1820, 1886, 1915, 1936, 2081, 2222, 2250, 2259, 2344, 2376, 2816, 2824, 2968, 2981, 3102, 3345, 3373, 3388, 3473, 3537, 3543, 5482, 5782, 6786, 7712 |

**Table 3.2.** Cleaned Distinct Work Box and Whisker Values

| Sample Size (n) | 1057 |
| --- | --- |
| Minimum | 1 |
| Maximum | 205 |
| Median | 19 |
| Lower Quartile (Q1) | 8 |
| Upper Quartile (Q3) | 53 |
| Interquartile Range | = Q3 - Q1  = 53 - 8  = 45 |
| Mean | 37.325449 |



**Figure 3.1.** Distinct Works Plot



**Figure 3.2.** Cleaned Distinct Works Plot

For the data of each author’s distinct works, we initially gathered a total of 1238 data for the sample size (n). Each corresponds to an author’s total published distinct works accessible on the Goodreads website. The dataset has a minimum value of 1 and a maximum value of 7712. From the dataset, we calculated the median to be 26 and a mean of 183.44830371567. We then calculate the lower quartile (Q1) and upper quartile (Q3) values. From the dataset, the lower quartile value (Q1) is 9 and the upper quartile value (Q3) is 88. With the lower and upper quartile values, we can then calculate the interquartile range by subtracting the lower quartile from the upper quartile. With that formula, we get the equation of 88-9 which results in 79 as the interquartile range. Following the interquartile range, we wanted to track the values that are considered outliers. We followed the lower and upper limit formulas. Where we multiply the interquartile range value by 1.5 and subtract it from the lower quartile value to get the lower limit or add it to the upper quartile value to get the upper limit. All the values that are greater than the upper limit or lesser than the lower limit are then considered outlier values. With the formula, we get an upper limit of 206.5 and a lower limit of -109.5.

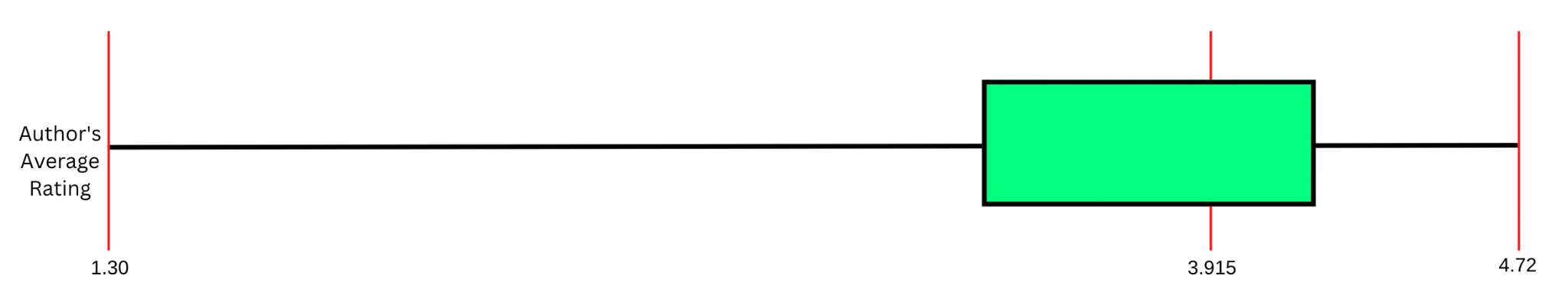
Once we figured out the upper and lower limit, we classified all the outlier values and removed them from our dataset to find the values of the data without the outliers. As the removal of outlier values can be considered data wrangling, we will refer to this new dataset as the ‘cleaned dataset’. This dataset contained 1057 total data for the sample size (n). We followed the same steps to find the minimum, maximum, median, mean, lower quartile, upper quartile, and interquartile range. The cleaned dataset has a minimum value of 1 and a maximum value of 205. The median is 19 and we calculated the mean to be 37.325449. We then calculated the lower quartile value to be 8, and the upper quartile value to be 53, which meant the interquartile range would be 53-8 which is 45.   
 After the values are found, we plot the values in a box and whisker plot to visualize the dataset. Figure 3.1 refers to the box and whisker plot of the whole distinct works data and Figure 3.2 refers to the box and whisker plot of the cleaned distinct works data.

**Table 3.3.** Author’s Average Rating Box and Whisker Values

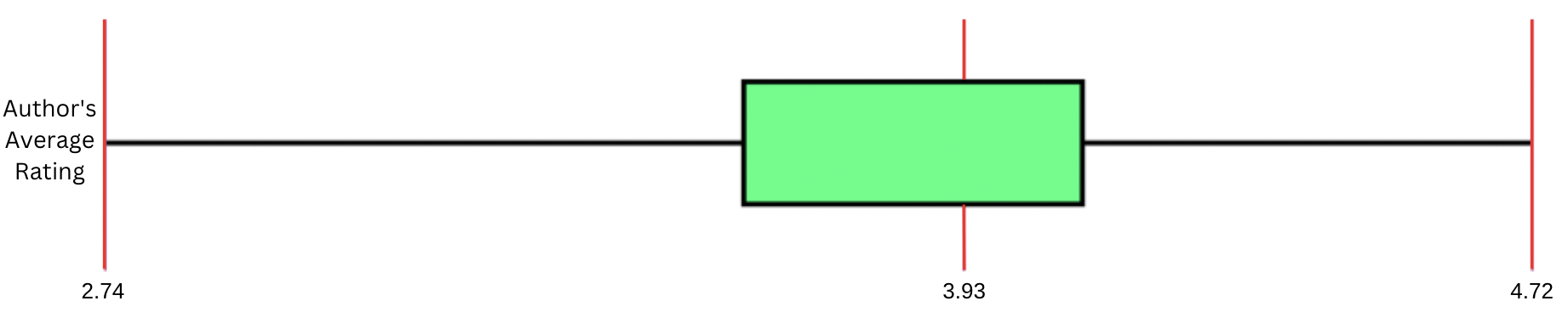
| Sample Size (n) | 1238 |
| --- | --- |
| Minimum | 1.30 |
| Maximum | 4.72 |
| Median | 3.915 |
| Lower Quartile (Q1) | 3.55 |
| Upper Quartile (Q3) | 4.09 |
| Interquartile Range (IQR) | = Q3 - Q1  = 4.09 - 3.55  = 0.54 |
| Mean | 3.7616397415186 |
| Outlier Values | Lower Limit = Q1 - 1.5×IQR = 2.74  Upper Limit = Q3 + 1.5×IQR = 4.9  Values above the upper limit or below the lower limit are considered outliers.  1.30, 1.48, 1.53, 1.62, 1.63, 1.64, 1.74, 1.86, 1.99, 2.06, 2.14, 2.17, 2.19, 2.25, 2.31, 2.34, 2.36, 2.39, 2.40, 2.40, 2.45, 2.47, 2.48, 2.49, 2.50, 2.53, 2.59, 2.60, 2.60, 2.60, 2.62, 2.62, 2.62, 2.64, 2.64, 2.65, 2.68, 2.68, 2.68, 2.69, 2.69, 2.70, 2.72 |

**Table 3.4.** Cleaned Author’s Average Rating Box and Whisker Values

| Sample Size (n) | 1195 |
| --- | --- |
| Minimum | 2.74 |
| Maximum | 4.72 |
| Median | 3.93 |
| Lower Quartile (Q1) | 3.62 |
| Upper Quartile (Q3) | 4.1 |
| Interquartile Range | = Q3 - Q1  = 4.1 - 3.62  = 0.48 |
| Mean | 3.813531 |



**Figure 3.3.** Author’s Average Rating Plot



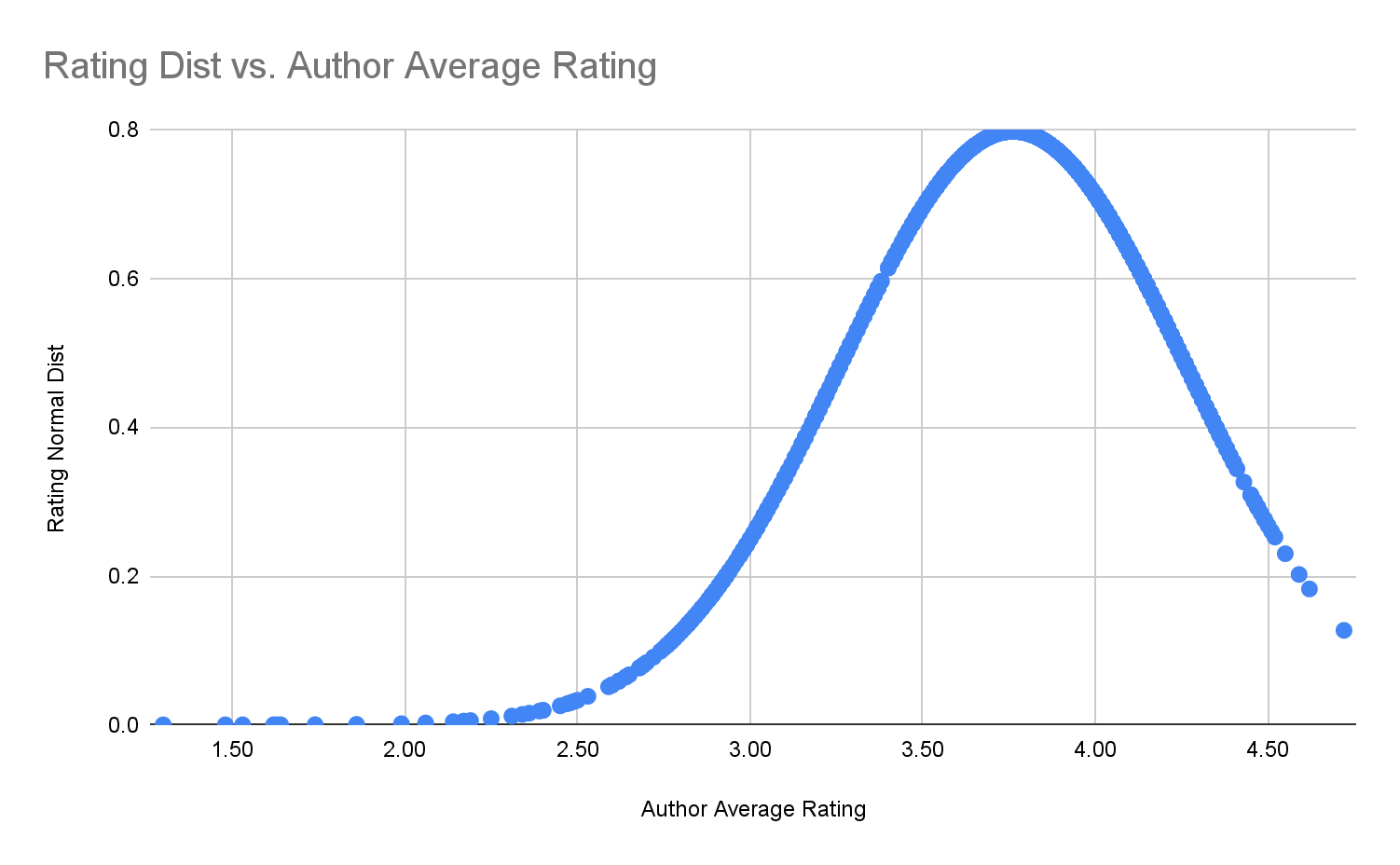
**Figure 3.4.** CleanedAuthor’s Average Rating Plot

For the data of each author’s average rating, we initially gathered a total of 1238 data for the sample size (n). Each corresponds to an author’s average rating as rated on the Goodreads website. The Goodreads website rating has a range of 1-5 with 1 being a very poor rating and 5 being the most excellent rating. The dataset has a minimum value of 1.30 and a maximum value of 4.72. From the dataset, we calculated the median to be 3.915 and a mean of 3.7616397415186. We then calculate the lower quartile (Q1) and upper quartile (Q3) values. From the dataset, the lower quartile value (Q1) is 3.55 and the upper quartile value (Q3) is 4.09. With the lower and upper quartile values, we can then calculate the interquartile range by subtracting the lower quartile from the upper quartile. With that formula, we get the equation of 4.09-3.55 which results in 0.54 as the interquartile range. Following the interquartile range, we wanted to track the values that are considered outliers. We followed the lower and upper limit formulas. Where we multiply the interquartile range value by 1.5 and subtract it from the lower quartile value to get the lower limit or add it to the upper quartile value to get the upper limit. All the values that are greater than the upper limit or lesser than the lower limit are then considered outlier values. With the formula, we get an upper limit of 4.9 and a lower limit of 2.74.

Once we figured out the upper and lower limit, we classified all the outlier values and removed them from our dataset to find the values of the data without the outliers. As the removal of outlier values can be considered data wrangling, we will refer to this new dataset as the ‘cleaned dataset’. This dataset contained 1195 total data for the sample size (n). We followed the same steps to find the minimum, maximum, median, mean, lower quartile, upper quartile, and interquartile range. The cleaned dataset has a minimum value of 2.74 and a maximum value of 4.72. The median is 3.93 and we calculated the mean to be 3.813531. We then calculated the lower quartile value to be 3.62, and the upper quartile value to be 4.1, which meant the interquartile range would be 4.1-3.62 which is 0.48.   
 After the values are found, we plot the values in a box and whisker plot to visualize the dataset. Figure 3.3 refers to the box and whisker plot of the whole author’s average rating data and Figure 3.3 refers to the box and whisker plot of the cleaned author’s average rating data.

1. **Distribution Graph**

After the box and whisker plot, we also decided to create a distribution graph to describe and show how the values of a variable are distributed.



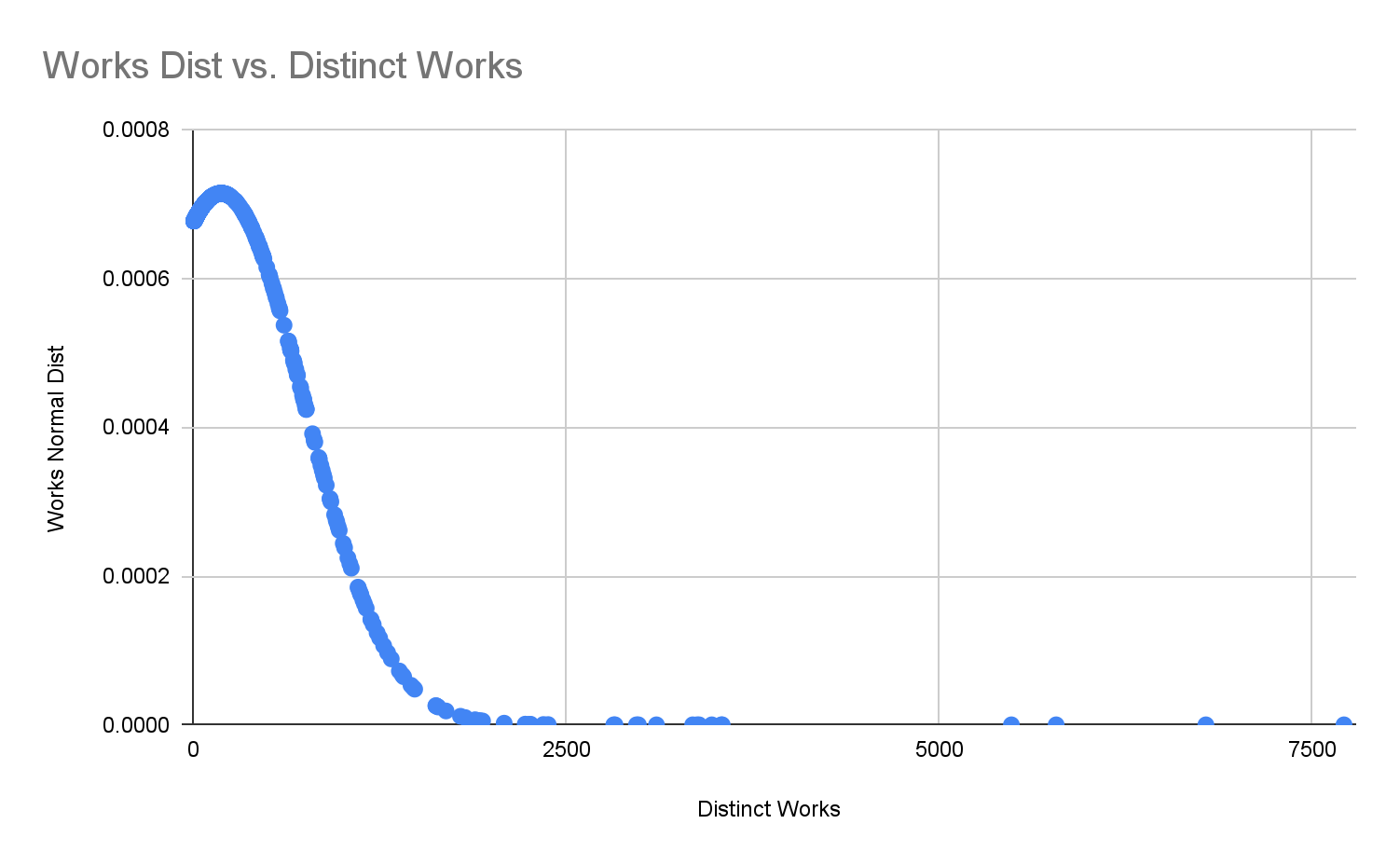
**Figure 3.5.** Author’s Average Rating Distribution Graph

We can see that for the author’s average rating data, the distribution graph the long tail is on the left side of the distribution, therefore it is skewed to the left, or in other words, it is negatively skewed. This shows that extreme values far from the peak are more common on the low (left) side compared to the high side (right).

Unlike in normal distribution, where the values for the mean, median, and mode are all similar, in an asymmetric distribution, those values are noticeably different. Checking the mean, median, and mode of the author’s average rating data we can see that the mean is less than the median, indicating that the mean underestimates the most common values because of one or more scores with relatively high values [4].

**Table 3.4.** Author’s Average Rating Mean, Median, and Mode

| Mean Rating | 3.76 |
| --- | --- |
| Median Rating | 3.92 |
| Mode Rating | 4 |



**Figure 3.6.** Author’s Distinct Works Distribution Graph

On the other hand, for the author’s distinct works data, we can see that the distribution graph’s long tail is on the right side of the distribution, making it a right-skewed or positively skewed distribution. This shows that extreme values far from the peak are more frequent in the high side (right) compared to the low side (left).

**Table 3.4.** Author’s Distinct Works Mean, Median, and Mode

| Mean Works | 183 |
| --- | --- |
| Median Works | 26 |
| Mode Works | 5 |

Checking the mean, median, and mode values for the author’s distinct works data shows us that the mean is greater than the median, indicating that the mean overestimates the most common values because of one or more scores with relatively low values [4].

# Result and Discussion

Before we discuss the results, we would like to note that we have decided to keep outliers when analyzing our dataset. The reasons why we did that are as follows:

* It is a natural part of the dataset [5].
* It was not a measurement or data entry error.

1. **Chi-Squared Test of Independence**

Chi-Squared test of independence can be used to determine whether or not we can reject the null hypothesis [6]. Below is the result for the author’s average rating and the number of distinct works by the author.

**Table 4.1.** Expected values table for average rating vs author works

| Author’s average rating | Number of distinct works | | | | | Total |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1-8 | 9-19 | 20-53 | 54-205 | >205 |  |
| 1-2 | 5.140243902 | 1.817447496 | 1.948303716 | 1.948303716 | 1.315831987 | 9 |
| 2-3 | 27.9184168 | 24.83844911 | 26.62681745 | 25.63327948 | 17.98303716 | 123 |
| 3-4 | 139.365105 | 123.9903069 | 132.917609 | 127.9579968 | 89.76898223 | 614 |
| 4-5 | 111.6736672 | 99.35379645 | 106.5072698 | 102.5331179 | 71.93214863 | 492 |
| Total | 281 | 250 | 268 | 258 | 181 | 1238 |

**Table 4.2.** Observation values table for average rating vs author works

| Author’s average rating | Number of distinct works | | | | | Total |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1-8 | 9-19 | 20-53 | 54-205 | >205 |  |
| 1-2 | 8 | 1 | 0 | 0 | 0 | 9 |
| 2-3 | 86 | 23 | 10 | 4 | 0 | 123 |
| 3-4 | 117 | 130 | 144 | 122 | 101 | 614 |
| 4-5 | 70 | 96 | 114 | 132 | 80 | 492 |
| Total | 281 | 250 | 268 | 258 | 181 | 1238 |

After calculating the expected values, we can then continue the chi-squared test using the formula;

**Table 4.3.** Expected values table for average rating vs author works

| Observation Values | Expected Values |  |  |  |
| --- | --- | --- | --- | --- |
| 8 | 5.140243902 | 2.859756098 | 8.178204938 | 1.591014958 |
| 86 | 27.9184168 | 58.0815832 | 3373.470307 | 120.8331522 |
| 117 | 139.365105 | -22.36510501 | 500.197922 | 3.589118826 |
| 70 | 111.6736672 | -41.67366721 | 1736.694538 | 15.55151346 |
| 1 | 1.817447496 | -0.817447496 | 0.6682204087 | 0.3676697182 |
| 23 | 24.83844911 | -1.838449111 | 3.379895135 | 0.1360751277 |
| 130 | 123.9903069 | 6.009693053 | 36.1164106 | 0.2912841454 |
| 96 | 99.35379645 | -3.353796446 | 11.2479506 | 0.11321108 |
| 0 | 1.948303716 | -1.948303716 | 3.795887368 | 1.948303716 |
| 10 | 26.62681745 | -16.62681745 | 276.4510584 | 10.38242963 |
| 144 | 132.917609 | 11.08239095 | 122.8193892 | 0.9240264711 |
| 114 | 106.5072698 | 7.49273021 | 56.141006 | 0.5271096152 |
| 0 | 1.948303716 | -1.948303716 | 3.795887368 | 1.948303716 |
| 4 | 25.63327948 | -21.63327948 | 467.9987812 | 18.25746805 |
| 122 | 127.9579968 | -5.957996769 | 35.4977255 | 0.2774170149 |
| 132 | 102.5331179 | 29.46688207 | 868.2971388 | 8.468455425 |
| 0 | 1.315831987 | -1.315831987 | 1.731413818 | 1.315831987 |
| 0 | 17.98303716 | -17.98303716 | 323.3896254 | 17.98303716 |
| 101 | 89.76898223 | 11.23101777 | 126.1357602 | 1.405115186 |
| 80 | 71.93214863 | 8.067851373 | 65.09022578 | 0.904883658 |
|  | | | | 206.8154212 |

We can get the meaning of the calculation is by seeing if the , chi-squared is greater or smaller than the . With our data, we have 4 rows and 5 columns, leaving us with 12 for our degrees of freedom and we chose 0.05 as our alpha value. With this, our chi-squared critical value is 28.3, which is smaller than our chi-squared calculated value of 206.82. Therefore, we reject the null hypothesis, meaning that we can accept the alternative hypothesis, where the number of distinct works affects the author’s average rating.

However, there is a possibility that there is a type 1 error, where our calculation results in a false positive, where there is a chance the result should not reject the null hypothesis but we rejected it [7]. This type 1 error can possibly happen if there is an observation bias, there is a possibility since we only have 1238 data points. There is also a possibility that a type 1 error can occur because we include outliers in our calculation. However, as explained in the previous section, the reason we include the outlier is that we do not want to mess with the order of data, and also they are all a natural part of the dataset, there is a minimum chance that the dataset is wrong. According to the p-value for the chi-squared independence test, there is a 0.00001 possibility that the null hypothesis should not be rejected, this shows minimizes error space for our calculation.

1. **Pearson Correlation Coefficient**

Pearson’s correlation is used to measure the correlation between two variables, to measure the strength of the correlation between them, indicated by the r-value [8]. R-value will be between 1 and -1. With 1 meaning that it has a strong positive correlation, and -1 as a strong positive correlation, the variables might be inversely proportional to each other. If the r-value is 0, that means it has no correlation. The closer it is the value to 0, the weaker the correlation that they have. Below is Pearson’s correlation calculation for our dataset.

| **Variable** | **Value** |
| --- | --- |
| Total author | 1238 |
| Sum of all ratings | 4656.91 |
| Sum of works | 227109 |
| Sum of ratings \* works | 898872.97 |
| Sum of ratings squared | 17826.6415 |
| Sum of works squared | 427093777 |

With the Pearson’s correlation formula;

From this calculation, we got 0.129 as our r-value. Therefore, based on Pearson’s correlation, this means that there is a positive correlation between the two variables. However, the r-value is so close to 0, that indicates that the correlation is very low. Testing this result by looking at its p-value of 0.05, to see the probability of the correlation to be equal to 0 or meaning no correlation. The p-value for this correlation test is 0.00001 with an r-score of 0.1291 and the number of data points is 1238. Therefore, we can validate that Pearson’s correlation coefficient test is proven to be right and our dataset is suitable for this test. We can use this result to support our hypothesis and also the chi-squared result.

1. **One-way Analysis of Variance (ANOVA)**

One-way analysis of variance, or ANOVA, is used to compare the means of three or more independent groups to ascertain whether there is a statistically significant difference between the corresponding population means [9]. The steps to perform the one-way ANOVA on this dataset are as follows:

1. Split the Author Distinct Works data into 5 groups

Since our intended data set has not been separated into groups, we create a table to store the values of the Author Average Rating based on which group from the Author Distinct Works they fall into. The ranges are determined from the box and whisker plots.

**Table 4.4.** Table of the first 5 Author Average Rating values of the table for the 5 groups derived from the Author Distinct Works

| Number of distinct works | | | | |
| --- | --- | --- | --- | --- |
| 1-8 | 9-19 | 20-53 | 54-205 | >205 |
| 1.3 | 1.62 | 2.47 | 2.76 | 3.02 |
| 1.48 | 2.06 | 2.76 | 2.78 | 3.05 |
| 1.53 | 2.17 | 2.77 | 2.89 | 3.46 |
| 1.63 | 2.19 | 2.88 | 2.99 | 3.46 |
| 1.64 | 2.31 | 2.89 | 3.04 | 3.5 |

1. Calculate the group means, the overall mean, and each group’s total count

After getting our 5 groups, we now have to calculate each group’s individual mean and also the overall mean for all the groups. In addition, we will also calculate the total count for each group.

**Table 4.5.** Table for the group means, overall mean, and group total count

| Number of distinct works | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | 1-8 | 9-19 | 20-53 | 54-205 | >205 |
| Group Count | 281 | 250 | 268 | 258 | 181 |
| Group Mean | 3.423772242 | 3.70916 | 3.854962687 | 3.945193798 | 3.958839779 |
| Overall Mean | 3.778385701 |  |  |  |  |

1. Calculate the regression sum of squares (SSR)

We will now calculate the regression sum of squares with the following formula:

1. Calculate the error sum of squares (SSE)

Next, we will calculate the error sum of squares with the following formula:

With

1. Calculate the total sum of squares (SST)

After finding both the SSR and SSE, we can now calculate the total sum of squares with the following formula:

+

1. Fill in the ANOVA table

Before we fill in the ANOVA table, we must find the following values:

With

**Table 4.6.** ANOVA table for the groups

| Source of Variation | Sum of Squares (SS) | Degrees of Freedom (df) | Mean Square (MS) | F |
| --- | --- | --- | --- | --- |
| Between | 50.83125001 | 4 | 12.7078125 | 60.6862380678331 |
| Error | 258.1925213 | 1233 | 0.2094018827 |  |
| Total | 309.3709402 | 1237 |  |  |

1. Interpreting the ANOVA results

Finally, to see whether the F-test statistic is statistically significant, we check an F-distribution table to compare the F-test statistic to the F-critical values there. We must also find the numerator degrees of freedom (df1) and the denominator degrees of freedom (df2).

Consulting the F-distribution table with , we can see that our F-critical value is . Since the F-test statistic is larger than the F-critical value we found, we can reject the null hypothesis; ergo, there is sufficient evidence to accept the alternative hypothesis, where the number of distinct works affects the author’s average rating.

1. **Kruskal-Wallis Test (One-Way ANOVA on Ranks)**

Since our distribution graph showed that our data sets both had skewed distributions, we also decided to use the Kruskal-Wallis test, also known as the one-way ANOVA on ranks. The steps to perform the Kruskal-Wallis test on this dataset are as follows:

1. Rank the data

We calculate the rank of one cell in comparison to the whole data set. We can do this by using the RANK.AVG() function in Google Sheets.

**Table 4.7.** Table of the first 5 values from the average rank table

| **Number of distinct works** | | | | |
| --- | --- | --- | --- | --- |
| **1-8** | **9-19** | **20-53** | **54-205** | **>205** |
| **A Ranks** | **B Ranks** | **C Ranks** | **D Ranks** | **E Ranks** |
| 1 | 4 | 22 | 47.5 | 140.5 |
| 2 | 10 | 47.5 | 55 | 153.5 |
| 3 | 12 | 51.5 | 90.5 | 274 |
| 5 | 13 | 84 | 130 | 274 |
| 6 | 15 | 90.5 | 148 | 291.5 |

1. Calculate the sample size, sum of ranks, and squared sum of ranks divided by sample size for each group

Before calculating the test statistic, we need to first calculate the sample size (n), sum of ranks (R), and squared sum of ranks divided by sample size for each group (R^2/n).

**Table 4.8.** Table for the group: sample size, sum of ranks, and squared sum of ranks divided by sample size

|  | **Number of Distinct Works** | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **1-8** | **9-19** | **20-53** | **54-205** | **>205** |
|  | **A Ranks** | **B Ranks** | **C Ranks** | **D Ranks** | **E Ranks** |
| n | 121545 | 146809 | 177548 | 190389 | 130650 |
| R | 281 | 250 | 268 | 258 | 181 |
| R^2 | 14773187025 | 21552882481 | 31523292304 | 36247971321 | 17069422500 |
| R^2/n | 52573619.31 | 86211529.92 | 117624225 | 140496012.9 | 94306201.66 |

1. Calculate the test statistic and find the critical chi-square value

After calculating the sample size, sum of ranks, and squared sum of ranks divided by sample size values for each group, we will now calculate the test statistic to find the critical chi-square value.

With

Now, we check a chi-squared distribution table to see the corresponding critical chi-square value, which is 9.488 at . Since the H-value is greater than the critical chi-square value, we can reject the null hypothesis. To check the probability that our conclusion is incorrect, we calculate the p-value with the H-value and degrees of freedom, we find that the p-value is < 0.00001. Therefore, there is a very small probability that the calculation is incorrect.

# Conclusion

In conclusion, our alternative hypothesis that an author’s number of distinct works does affect the author’s average rating was right. From the dataset of a list of 1238 authors' average rating and the number of works they made, we found that the chi-squared test of independent value that we calculated is greater than the critical value with degree of freedom of 12 and 0.05 alpha, 206.8 > 28.3. In addition, the results of the one-way ANOVA test and Kruskal-Wallis test we performed also support the chi-squared test’s results. The ANOVA test yielded an F-test statistic that is greater than the F-critical value, with degrees of freedom (4, 1233) and 0.05 alpha, 50.99 > 2.37. Meanwhile, the Kruskal-Wallis test also resulted in an H-test statistic that is greater than the critical chi-square value, with degrees of freedom 4 and 0.05 alpha, 125.89 > 2.37. We conclude that we reject the null hypothesis and automatically accept the alternative hypothesis.

We also use pearson’s correlation to our dataset that results with r-value of 0.129. Concluding the positive low correlation of the two variables. This supports the chi-squared test of independence, that there is an effect between the number of works and author’s average rating. We can conclude that both variables have a proportional correlation, however it is a weak one. Meaning that it has a little impact on the rating, considering there are other variables that were taken to a matter to calculate the author’s ranking.

# References

[1] “About Goodreads,” *Goodreads.com*, 2019. https://www.goodreads.com/about/us

[2] C. S. Ross, “Making Choices,” *The Acquisitions Librarian*, vol. 13, no. 25, pp. 5–21, Aug. 2000, doi: 10.1300/j101v13n25\_02.

[3] K. Saarinen and P. Vakkari, “A sign of a good book: readers’ methods of accessing fiction in the public library,” *Journal of Documentation*, vol. 69, no. 5, pp. 736–754, Sep. 2013, doi: 10.1108/jd-04-2012-0041.

[4] R. S. Witte and J. S. Witte, “Describing Data with Averages,” in *Statistics*, 11th ed., Hoboken, NJ: Wiley, 2017, pp. 53–56.

[5] J. Frost, “Guidelines for Removing and Handling Outliers in Data - Statistics By Jim,” Statistics By Jim, Oct. 23, 2019. <https://statisticsbyjim.com/basics/remove-outliers/>

[6] M. L. McHugh, “The chi-square test of Independence,” *Biochemia Medica*, vol. 23, no. 2, pp. 143–149, Jun. 2013.

[7] A. Banerjee, U. B. Chitnis, S. L. Jadhav, J. S. Bhawalkar, and S. Chaudhury, “Hypothesis testing, type I and type II errors,” *Industrial Psychiatry Journal*, vol. 18, no. 2, pp. 127–131, 2009.

[8] P. Schober, C. Boer, and L. A. Schwarte, “Correlation coefficients,” *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.

[9] Zach, “How to perform a one-way anova by hand,” *Statology*, 03-Mar-2020. [Online]. Available: https://www.statology.org/one-way-anova-by-hand/. [Accessed: 07-Jan-2023].