Data Analysis

1st Mehmet Emre Aydın
Galatasaray University
Istanbul, Turkey
emre-aydin-079@hotmail.com

2ndAlparslan Aslan
Galatasaray University
Istanbul, Turkey
alparslanaslan.1997@gmail.com

***Abstract*—In this study, we examined the best selling 271 programming book dataset in America. We have presented several hypotheses to reveal the popularity, number of pages, and the relationships between the books. To test these hypotheses, we applied a t-test, z-test, regression analysis, and covariance operations on the dataset and obtained numerical results. We evaluated the results by supporting them with seven ads we found.**

***Index Terms*—data analysis, regression analysis, t-test,correlation**

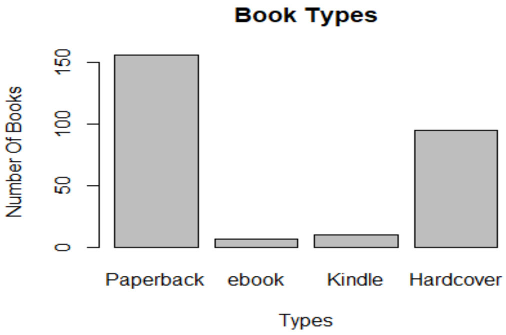
I. Introduction

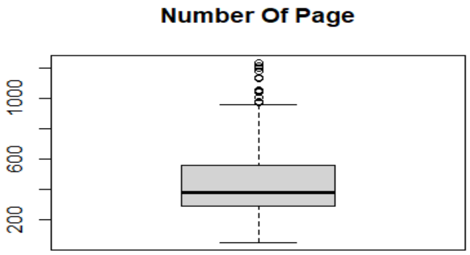
We will be working on the best programming books dataset written in English. There are 271 instances in our dataset. There are 7 features of our data. The rating feature represents the average of the points users give to the book. The Review feature represents the number of comments people make on the book, the Book Title feature is the title of the book, The description is a brief introduction of the book's content, the Number Of Page is the number of pages of the book, Price is the equivalent of the average sales prices of the book on 5 different websites and Type represents the publication type of the book. Our data has 5 different types. These are the e-book, Hardcover, Kindle Edition, Paperback. Review, Rating, Number Of Page, and Price are Quantitative and Interval features. Type is a Nominal feature.

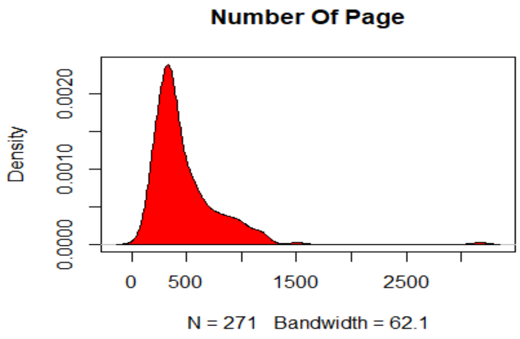
We have shown below the distribution charts and standard deviation, mean and median values for a clearer understanding of the data.

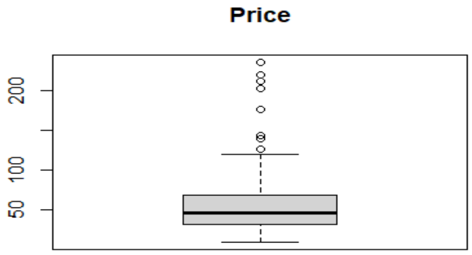
- Does the Price increase as the Number Of Pages increases?
- Is there a link between Reviews and Price?
- Does Rating increase when Number Of Page increases? We expect that there will be more criticism in books with more pages.
- Does the book types being Hardcover or Paperback affect Rating?
- Do Hardcover books have more pages than Paperback?
- Do Kindle Edition books get more ratings than ebook?
- Ebooks have more Price than Kindle Edition?

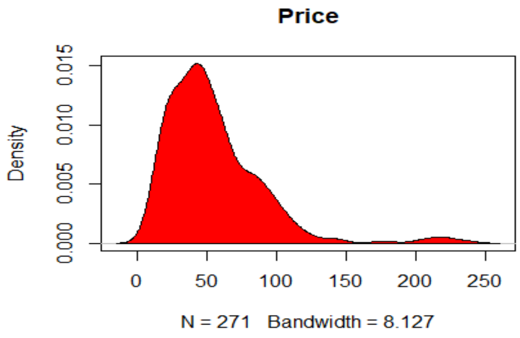
	Rating	Reviews	Number Of Page	Price
Mode	4.15	3	288	51,473
Mean	4.0674	185.557	475,077	54,541
Median	4.1	35	384	46,317
Variance	0.084	304,839	93.726	1275,31
Standart Deviation	0.290	552,122	306,147	35,751

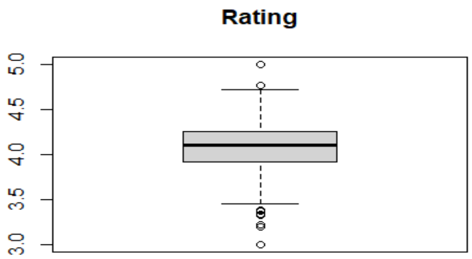


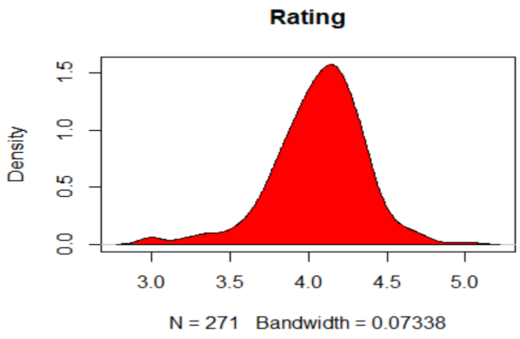


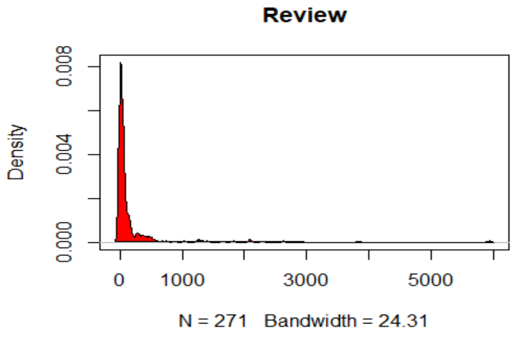












II. Hypotesis

- $P1 : \mu > 0,7$
  $P0 : \mu < 0,7$
  μ = "Result Of Price and Number of Page's Correlation "
- $P1 : \mu > 0,7$
  $P0 : \mu < 0,7$
  μ = "Result of Price and Review's Correlation "
- $P1 : \mu > 0,7$
  $P0 : \mu < 0,7$
  μ = "Result of Rating and Number of Page's Correlation "
- $P1 : \mu < -1,28155$
  $P0 : \mu > -1,28155$
  μ = "T-Score of Hardcover and Paperback's Rating "
- $P1 : \mu > 1,28155$
  $P0 : \mu < 1,28155$
  μ = "T-Score of Hardcover and Paperback's Number Of Page "
- -$P1 : \mu < -1,65$
  $P0 : \mu > -1,65$
  μ = "Z-Score of Kindle Edition's Rating "
- $P1 : \mu < -1,65$
  $P0 : \mu > -1,65$
  μ = "Z-Score of Ebook's Price "

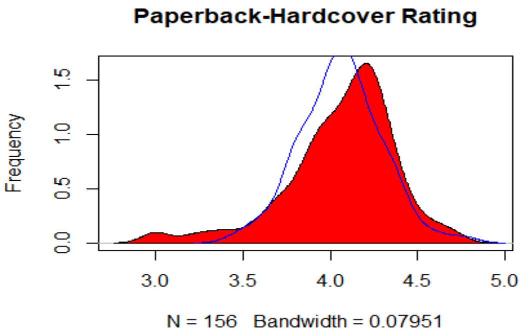
III. Methods

A. T-Test

- Paperback-Hardcover Rating t score=-0,11337

As seen above, the T-Score on the paperback and hardcover's rating is greater than -1.28155. This means that the

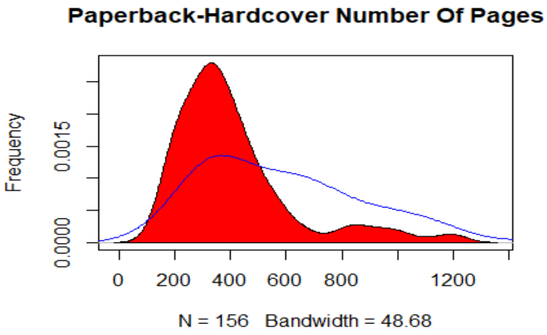
Null hypothesis is correct. So whether the book is Hardcover or Paperback does not affect the Rating. Below is a graph showing this situation.



The chart above is our rating chart for Hardcover and Paperback. The blue color represents Hardcover and the red color represents Paperback. As we can see in the graph, 2 data share a common area at many points. This shows that Hardcover and Paperback do not affect the rating. Below is a graph showing this situation.

- Paperback-Hardcover Number Of Page t score=4,7481

As seen above, the T-Score on the Number Of Page of Paperback and Hardcover is greater than 1.65. This shows that the Alternative Hypothesis is correct. In other words, the fact that the book is Hardcover or Paperback affects Number Of Page.



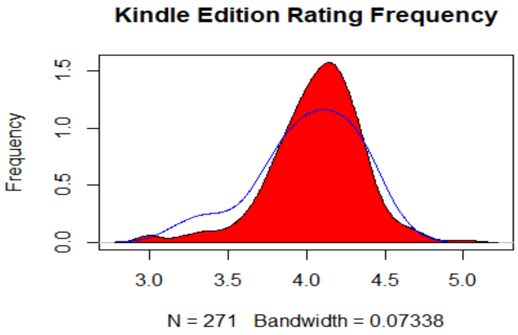
The chart above is our Hardcover and Paperback's Number Of Page chart. The blue color represents Hardcover and the red color represents Paperback. As we can see in the graph, the 2 data have many differences. We can say that the Hardcover has more Number Of Pages. This shows that it affects Hardcover and Paperback Number Of Page.

B. Z-Test

$$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} \tag{1}$$

- Kindle Edition Rating z score=-0,51803

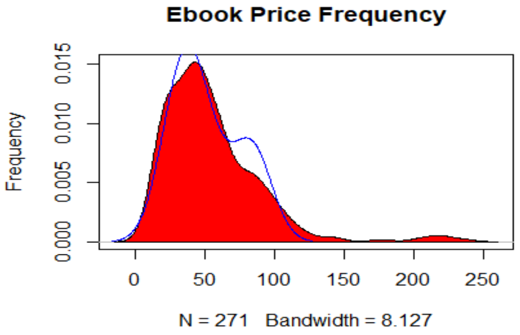
As seen above, Kindle Edition's Z-Score on Rating is greater than -1.65. So we see that our Null hypothesis is correct. This is not more than the rating of the Kindle Edition to us.



The graph above shows the general population in red. Blue shows the distribution of the Kindle Edition. It seems that these data are different at many points. In other words, as we can see in the graph, the rating of the Kindle Edition is not higher than the general population.

- Ebook Price z score=-0,36045

As seen above, the Ebook's T-Score on Price is greater than -1.65. So we see that our Null hypothesis is correct. This is not more than the price of the Ebook to us.



The graph above shows the general population in red. Blue shows the distribution of the Ebook. It seems that these data are different at many points. So as we can see in the graph, the Ebook's Price is not more than the general population.

C. Correlation and Covariance

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}} \tag{2}$$

- Number Of Page - Price cor=0,6950663

The fact that the correlation result is very close to 0.7 tells us that there is a relationship between the Number Of Pages and Price. However, to reach a more detailed conclusion, its graph should be examined.

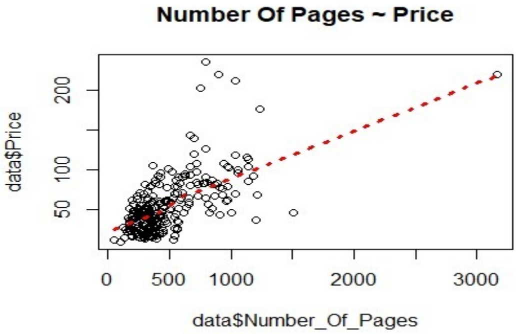
- Number Of Page - Rating cor=0,132288

We can say that there is no relationship between these two features by looking at the very small correlation result.

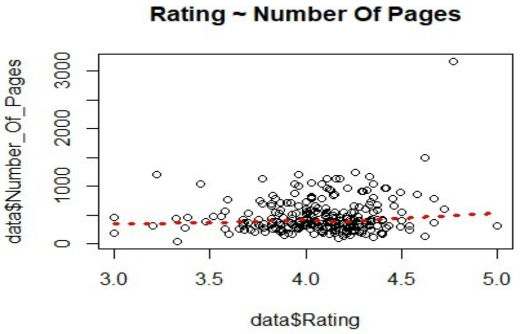
- Reviews - Price cor=0,2548988

By looking at the very small correlation result, we can say that there is no relationship between Reviews and Price features.

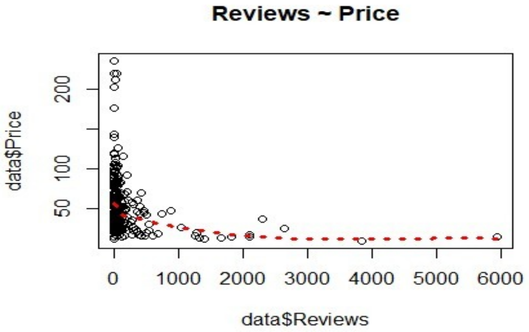
D. Regression Analysis



It is clearly seen in the chart above that as the number of pages of the book increases, so does the price. Based on this graphic, we can say that there is a relationship between the two features.



When we examine the graph above, the slope of the linear function found as a result of the regression analysis is close to zero. This tells us that the Rating and Number Of Pages features are irrelevant.



When the graph above is examined, we cannot obtain a simple linear function.There is a graph similar to a

$$y = e^{1/x}$$

function chart. We get such a graphic due to the large number of books that do not have comments. When we ignore this, we can easily say that there is no relationship between Reviews and Price.

IV. Conclusion

We used data from the best selling computer programming books in America. We prepared 7 different questions on this data that we were curious about. We reviewed these questions using data analysis methods and prepared some technical questions. We obtained some numerical data by applying a T-test, Z-test, Correlation analysis and regression analysis methods on the data. Based on the numerical data and graphics we have obtained, we found that some of the 7 questions we have prepared are correct and some are wrong. The results we found are as follows.

- Number Of Page increases when Price increases.
- There is no relationship between Reviews and Price.
- When Number Of Page increases, the Rating does not increase.
- Whether the book is Paperback or Hardcover does not affect the Rating.
- Hardcover books have more Number Of Page than Paperback.
- Kindle Edition type books do not have more Rating.
- The price of ebook type books is not more.