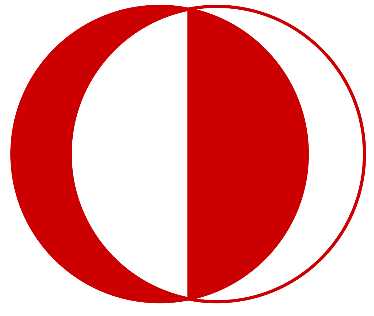
**ÖYKÜCÜ BOOKSTORE SALES INFORMATION**



A FINAL PROJECT REPORT SUBMITTED

IN FULFILMENT OF THE REQUIREMENTS FOR THE COURSE

STAT 112

INTRODUCTION TO DATA PROCESSING AND VISUALIZATION

DEPARTMENT OF STATISTICS OF

MIDDLE EAST TECHNICAL UNIVERSITY

BY

BUSE ERDEM / 2561231

MERVE AYDEMİR / 2561058

ZEYNEP KANAR / 2561306

FİTNAT KOÇ / 2450542

BERFİN AKDEMİR / 2614410

January 2023

**ABSTRACT**

In the past, to buy books, people would go out and go around bookstores, bibliopoles, garage sales, etc. As the internet paved its way into people’s lives, it also changed our shopping habits, including book sales. Öykücü Bookstore kept abreast of the revolution of online shopping. In this analysis, after cleaning the data, with the sales information of the Öykücü Bookstore, the relationships between the categories of the books and their prices, ratings, and clicking rates were examined concerning the years and publishers using descriptive statistics and exploratory data analysis techniques.

From the analysis, the following were found:

* The novel and history books are rated higher.
* There is a linear positive relationship between price and clicking rate
* The second-hand books are cheaper.
* Novel and history books have the higher price range, while science books prices are accumulated around 80 Turkish Liras.
* Doğan Publishing has generally higher prices than Can Publishing.
* The highest and lowest clicking rates of books are dominated by the novel category.
* The website clicking rate increases over the years.
* The number of pages does not have a considerable effect on the shipping fee.

1. INSTRUCTION

This analysis mainly examined the prices and the ratings of the books at the Öykücü Book Store with respect to the other sales information. The dataset contains the sales information for 119 books sold at the Öykücü Book Store. The dataset has 10 different variables:

* Price: the price of the book
* Currency Unit: the currency unit of the price (tl)
* Publisher: the publisher of the book
* Type: the type of the book
* No of Pages: the number of pages in the book
* Rating: the average rating of the book on the website
* Clicking Rate: the average clicking rate of the book on the website
* Shipping Fee: the shipping fee of the book if it is ordered
* Year: the published year of the book
* Second Hand: the book is used before or not(yes: second hand, no: newly published)

# Data description

This dataset includes 124 rows and 10 columns. Here, the most used variables by price are clicking rate, type, secondhand, and publisher.

* 1. Research questions

Here, 8 research questions were established by looking at the relationship between the price and rating of the books with other variables.

* Which type of books are liked more?
* How does the price affect the clicking rate of a book?
* How does the book being second-hand affect the price of the book?
* How does the category of the books affect the price range and density of the prices?
* How does the publisher of the books affect the price range and the density of the prices?
* How does the clicking rate of the website change over the years?
* Which type of books are looked at more by consumers?
* Do the heavier books cost more for shipping?

1. Data Tidying and Cleaning Steps and Exploratory Data Analysis
   1. Data Tidying and Cleaning Steps
2. The pandas, numpy, seaborn and matplotlib libraries were imported, data was read into the notebook file with pandas.
3. The first and last 5 lines of the data were examined for the existence of header, separation errors and NA values with \_head() and \_tail() functions.
4. The data types and value counts of the variables were examined with \_info() function.
5. All of the column names were reshaped to correct mispells and practicality with \_title() and \_rename() and \_title() functions.
6. The descriptive statistics of the variables were examined with the \_describe() function in order to observe outliers. There were outliers observed for price, shipping fee and rate variables with the values of 500, 210 and ≈13 respectively.
7. The NA values that were found for categorical variables such as type, Publisher and currency unit were filled with mode.
8. The data was checked for duplicate values with \_duplicated().sum() function.
9. After observing the data with \_valuecounts() typos were observed such as “DOgan”, “YeS” , “tl”.
10. The typos were fixed to the same format wtih .loc[] function.
11. The data was observed for null values and their counts, there were null values for price, number of pages, rating, clicking rate and shipping fee.
12. The numeric outlier values that was found at the previous steps (number of pages, rating, shipping fee) were replaced with the mean of the variable without outliers with indexing of lists.
13. The null values of the numeric variables were filled with the mean that was computed without outliers with \_fillna() function.
14. Then the data was examined again for the existence of the outliers and null values.
15. The year variable was corrected to ‘date’ data type from ‘object’.
16. The data was examined for the last time to ensure it is clean.
17. Clean data was exported to a new excel file.
    1. Exploratory Data Analysis
18. Which type of books are liked more?

Chart, box and whisker chart

Description automatically generated

Book type is a categorical, and rating is a numerical variable. To present this binary variables box-plot is very efficient . As represented on the box-plot median value is similar for all three of the variables with a value between 5-6. Even though the average rate is close for each of them the spread varies. The novels has the widest range of ratings followed by history books , as oppose to this, science books have a considerable narrow range. With the difference of the range it is also seen that even though the science books were not rated higher than ≈8, it was also not rated above ≈3; unlike history and novels. Even though all three of the plots might look symmetrically distributed , science plot has a slight right skew. It is also seen that the lowest rated books are at the novel category. In conclusion, novels are the most liked and disliked type of books in our data.

1. How does the price affect the clicking rate of a book?

Chart, scatter chart

Description automatically generated

One of the best methods to determine the relationship between two numeric variables is to use a scatter plot. We used the scatter plot to show whether there is a relationship between clicking rate and price. According to the graph, we can say that there is a weak linear relationship between the variables. Observing the correlation coefficient which is r ≈0.548 , A positive linear correlation indicates that, although both variables tend to go up in response to one another, the relationship is not very strong. Nevertheless, expensive books have a higher click rate than cheap books. We can say that this is because people do more research when buying expensive books. Since cheap books are more readily available, people may have felt the need to do less price research

1. How does the book being second hand affect the book price?

Chart, bar chart

Description automatically generated

The bar graph is one of the most appropriate methods to observe whether the second hand books have an effect on the prices. As seen in the graph , second hand books costs less with the value of ≈60 Turkish Liras on average.

1. How does the category of the books affect the price range and the density of the prices?

Chart, radar chart

Description automatically generated

Violin plots are beneficial in presenting multiple variables with their density and spread. That is why we use violin plots to answer this question. According to the graph, we can say that the prices of history and novel books spread more widely. That is to say, these types of books are sold at a wide range of prices. For example, finding affordable and expensive history and novel books is possible. Observing the science books prices, they have considerably narrow range with the price being accumulated around 80 Turkish Liras. So there is no big difference between the book prices. It is seen that the median values ​​of different types of book prices are similar. The width of the plots also provides information about the distribution. We can see that in the regions where the width is larger, more books are sold at the corresponding prices, and in the regions where the width is narrower, fewer books are sold at the corresponding prices. Hence, novel and history books are priced on a wide range of prices, while science books costs around similar values.

1. How does the publisher of the books affect the price range and the density of the prices?

Chart, radar chart

Description automatically generated

The violin plot can also be used to analyze the effect of publishers on book prices. The graph shows the effect of three different publishers on book prices. According to the graph, while we can comment on Can and Dogan publishing, we cannot include ODTÜ in these comparisons since there is only one data on ODTÜ publishing in our dataset. When we look at Can and Doğan, we see that the book prices of Doğan publishing spread more regularly than Can publishing. In addition, the median value of Doğan publishing is also larger than Can publishing with a value of ≈90 . When the interquartile range is analyzed, we see that Can publishing is larger with a value of ≈10. Therefore, it can be said that Can Publishing's pricing is more diversified. In addition, the median value is closer to Q3, indicating that even though Can publishing house sells very cheap books, the media is closer to expensive books. Thus, it is possible to say that Doğan Publisher books are generally more expensive than Can.

1. Which types of books are looked at more by comsumers?

Chart, box and whisker chart

Description automatically generated

To display clicking rate and type relationship, box-plot was used. Observing the spread we see that novels has the widest range for clicking rates, followed by history books. It is also seen that even though it has a narrow range for clicking rates, science books have the most regular spread, observing that beside the outliers, the lowest clicking rate starts at ≈5. Looking at the box-plots it is seen that science books clicking rate might be right skewed with their mode around ≈5.1, novels are slightly left skewed with the mode clicking rate of ≈5, and history books have a possibly right skewed distribution for clicking rates with a mode around ≈ 5.2. Examining the outliers, it is possible to say that, in contrary to their spread, there are history books with prominent popularity, and there are unregarded science books. (…) In concluison novels takes both highest and lowest curiosity by the consumer.

1. How does the clicking rate of the website changed over the years?

Chart, line chart

Description automatically generated

Since we have time as a variable, the line plot is the best choice to visualize. According to this plot, we can observe how the average clicking rate changes over the years. Until 2020, there was an increase in the average clicking rate, while we see that there was a significant decrease in 2021 and a rapid increase after 2021. From 2015 to 2020, the increase in the clicking rate may be due to the popularization of online shopping over time and the increase in accessibility to technology over time. In addition, considering that the COVID-19 virus started in the 2020-2021 period, the reason for the decrease between these years can be interpreted as the fact that people followed the news and social media more instead of reading books during this period.

8. Does heavier books cost more for shipping?

Chart, scatter chart

Description automatically generated

Considering economical and administrative influences it would be expected that for a small bookstore to conduct the shipping fees according to the weight of the books. When we want to analyze the effect of book weight on shipping fees, we can compare these two numeric variables using a scatter plot. Since the book weight will increase as the number of pages increases, we cannot see a strong positive or negative linear relationship between the points when we also compute the correlation coefficient which is r ≈ -0.102. Therefore, contrary to the expected outcome, the number of pages in a book does not have a strong effect on shipping fees.