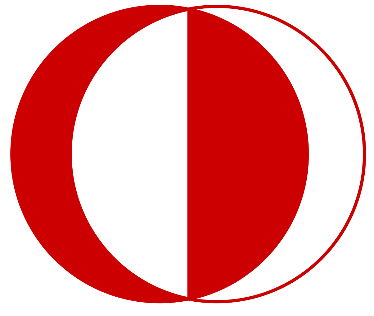
**ÖYKÜCÜ BOOKSTORE SALES INFORMATION**



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

Name

January 2023

**1.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 positive linear relationship between price and clicking rate
* Second-hand books are cheaper.
* Novel and history books have a higher price range, while science books prices are accumulated around 80 Turkish Liras.
* Doğan Publishing generally has higher prices than Can Publishing.
* The highest and lowest clicking rates of books are dominated by the novel category.
* The website clicking rate has increased over the years.
* The number of pages has little effect on the shipping fee.

**2.INTRODUCTION:**

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)

**2.1 Data Description:**

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

**2.2 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 changed over the years?
* Which type of books are looked at more by consumers?
* Do the heavier books cost more for shipping?

**3.DATA TIDYING AND CLEANING STEPS:**

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

**4. EXPLORATORY DATA ANALYSIS:**

In this Project, we conducted 8 research questions.

**1-) Which type of books are liked more?**

Book type is categorical, and rating is a numerical variable. To present these binary variables, a 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 have the broadest range of ratings, followed by history books. In opposition to this, science books have a considerably narrow range. With the difference in 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, the science plot has a slight right skew. It is also seen that the lowest-rated books are in the novel category. In conclusion, our data show that novels are the most liked and disliked type of books.Chart, box and whisker chart

Description automatically generated

Chart, scatter chart

Description automatically generated

**2-) How does the price affect the clicking rate of a book?**

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, there is a weak linear relationship between the variables. Observing the correlation coefficient, 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.

**3-) 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 second-hand books affect the prices. The graph shows that second-hand books cost less, with a value of ≈60 Turkish Liras on average.

On the other hand, the average price of non-second-hand books is approximately 80 TL. Considering the price difference between second-hand and non-second-hand books, the fact that a book is second-hand reduces the price value of the book.

**4-) 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, the prices of history and novel books spread more widely. 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 a considerably narrow range, with the price at around 80 Turkish Liras. So, there is no significant 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 more extensive, 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 at a wide range of prices, while science books cost around similar values.

**5-) 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 analyse 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 analysed, we see that Can publishing is more significant 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.

**6-) Which types of books are looked at more by consumers?**Chart, box and whisker chart

Description automatically generated

A box plot was used to display the clicking rate and type relationship. Observing the spread, we see that novels have 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 besides 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, contrary to their spread, there are history books with prominent popularity, and there are unregarded science books. In conclusion, novels take the highest and lowest curiosity of the consumers.

**7-) How does the clicking rate of the website change over the years?**

**8-) Does the heavier books costs more for shipping?**Chart, line chart

Description automatically generated

Considering economic and administrative influences, it would be expected for a small bookstore to conduct the shipping fees according to the weight of the books. When we want to analyse 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 compute the correlation coefficient, r ≈ -0.102. Therefore, contrary to the expected outcome, the number of pages in a book does not strongly affect the shipping fee.Chart, scatter chart

Description automatically generated

**5-) CONCLUSION:**

To conclude, this study focused on 3 main subjects: the range and impact of price, the factors which affect the clicking rate, and the general impact on the category of the books. Several statistical methods were used on the data such as correlation coefficient, measures of central tendency and variability, and five-number summary. In general, novels were the center of curiosity, likability, and affordability, while also involving the most expensive, disfavored, unpopular books in themselves. Consumers have looked at expensive books more. Examining the spread of price, it was seen that science books have the most consistent range among the categories, it was also seen that the Doğan publishing books generally cost more and second-hand books are more budget-friendly. Going forward with costs, it was also seen that the number of pages of a book does not have a significant effect on the shipping fee. Lastly, it was seen that online shopping had its effect on written goods as well as it did on other consumers goods also like the internet did, online shopping for books was also possibly affected by the pandemic with having an eye-catching change in clicking rates with the outbreak.