My Data Literacy Project
(Replace this with your Project Title)

Nico Rinck [*1] **Adriano Polzer** [*2] **Robin Allgeier** [*3] **Jonas Mahr** [*4] **Jannik Rombach** [*5]

Abstract

Put your abstract here. Abstracts typically start with a sentence motivating why the subject is interesting. Then mention the data, methodology or methods you are working with, and describe results.

1. Introduction

Public libraries play an important role in providing people with access to knowledge, education, and entertainment. To maintain efficient operations and ensure equitable resource availability, libraries must understand how their users interact with the collection. Analyzing borrowing behavior provides valuable insights into usage patterns, user preferences, and operational challenges, enabling data-driven decision-making to improve service quality. Understanding borrowing behavior encompasses multiple dimensions: which types of media are most frequently borrowed, how different user groups interact with the library, how loan durations and extension patterns vary across media types and user categories, and temporal trends in borrowing activity. One particularly important aspect of borrowing behavior is the timely return of borrowed items. Late returns can lead to reduced availability of popular materials, diminished satisfaction among users waiting for specific items, and increased administrative burdens for library staff. Identifying factors associated with late returns enables libraries to adapt their policies, such as loan duration limits, extension rules, or targeted reminder systems, to improve resource circulation and user satisfaction. This study analyzes borrowing records from a public library. The primary objective is to characterize borrowing behavior across multiple dimensions and to identify patterns that contribute to our understanding of library usage. Specifically, this study addresses the following questions: *How do borrowing patterns vary across media types, user categories, and time periods? What observable characteristics of a loan are associated with late returns?* The analysis investigates temporal patterns, in borrowing activity, and examines the relationships between late returns and several factors.

Our analysis reveals several clear patterns: . . . (To Do)

2. Data and Methods

2.1. Data Collection and Description

This study investigates the borrowing behavior of users at the Tübingen City Library, as well as the frequency of late returns. The dataset was provided directly by the Tübingen City Library and covers borrowing records from 2019 to 2025. It encompasses over 2.4 million individual loan transactions and includes the following key variables: borrowing and return timestamps, media types, user categories, number of extensions, anonymized user identification numbers, and late return flags. Each record documents a complete transaction from initial borrowing to return. Domain knowledge was gathered through a personal interview with a staff member of the Tübingen City Library, providing contextual insights into the data collection processes, library operations, and data quality practices. Additional information was obtained through ongoing email communication with the library staff, which helped to clarify data inconsistencies and provide domain expertise for informed analysis decisions.

2.2. Data Quality Assessment: Sanity Checks

Prior to conducting any analysis, a comprehensive set of data quality checks was performed to validate internal consistency and identify potential data quality issues. These sanity checks were designed to detect anomalies without filtering or removing data at first, and gather information, regarding cleaning the data to ensure robust analysis afterwards. The following specific checks were implemented:

[*]Equal contribution [1]Matrikelnummer 12345678, MSc Computer Science [2]Matrikelnummer 12345678, MSc Computer Science [3]Matrikelnummer 12345678, MSc Computer Science [4]Matrikelnummer 12345678, MSc Computer Science [5]Matrikelnummer 7317181, MSc Computer Science. Correspondence to: Initials1 <first1.last1@uni-tuebingen.de>, Initials2 <first2.last2@uni-tuebingen.de>, Initials3 <first3.last3@uni-tuebingen.de>, Initials4 <first4.last4@uni-tuebingen.de>, Initials5 <jannik.rombach@uni-tuebingen.de>.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2025/26 (Module ML4201). Style template based on the ICML style files 2025.

2.2.1. Missing Values

First of all, the overall of missing values in the dataset was assessed. The proportion of missing values in each column was calculated to identify any fields with significant gaps that could impact the analysis. Depending on the extent and nature of the missing valus, they were excluded from certain analyses. Return timestamps are missing in roughly two percent of rows, which is expected for currently borrowed and unreturned or lost items and therefore retained. The same rate applies to the derived duration column because it depends on return timestamps. A missing rate of about 6.7% was found in the user ID colum, which will be handled accordingly in the analysis. All other colums, which are relevant for our analyses, have a missing rate below 0.1% and are therefore considered complete.

2.2.2. Timestamp and Duration Consistency

The integrity of temporal data was verified by identifying logical inconsistencies, such as instances where return dates preceded borrowing dates. These checks are critical since all downstream analyses depend on accurate temporal information. Further checks on the duration value revealed that it is calculated correctly from the timestamps and does not contain any inconsistencies.

2.2.3. Late Return Consistency

The consistency between late return flags and the reported number of days late was validated. No Entries were identified where the late flag indicated "not late" but the days-late counter was positive, and conversely, records marked as late but with zero or missing days-late values. Additionally, implausible values in the extensions column (negative extensions or more than six extensions) were checked. There are a view cases where the number of extentions is higher than six, which should not be possible according to the library's policies. But because of the low number of affected rows (0.003%), these where not further investigated.

2.2.4. Duplicate Analysis

Various forms of duplicates were examined. There are no exact duplicates in the dataset. However, some entries show the the exact same borrowing timestamp. These can be explanied by the fact, that users can borrow at multiple stations at the same time. Also it is possilbe for user to borrow multiple items in one transaction, resulting in identical borrowing timestamps for different items, but all these occurances limit to the same user ID and to a maximum of seven items/entries per transaction. Only one case was found where a diffetent user borrowed and returned at the exact same times, which is considered a rare but possible event and therefore retained.

2.3. Analysis Approach

For the analysis of borrowing behavior and late returns, descriptive statistical methods were employed to identify patterns and trends in user behavior. These methods were chosen to provide a comprehensive understanding of borrowing dynamics and to isolate factors associated with late returns. The analysis focuses on temporal patterns, user segments, media types, and their relationships with return timeliness.

3. Results

In this section outline your results. At this point, you are just stating the outcome of your analysis. You can highlight important aspects ("we observe a significantly higher value of x over y"), but leave interpretation and opinion to the next section. This section absoultely *must* include at least two figures.

4. Discussion & Conclusion

Use this section to briefly summarize the entire text. Highlight limitations and problems, but also make clear statements where they are possible and supported by the analysis.

Contribution Statement

Explain here, in one sentence per person, what each group member contributed. For example, you could write: Max Mustermann collected and prepared data. Gabi Musterfrau and John Doe performed the data analysis. Jane Doe produced visualizations. All authors will jointly wrote the text of the report. Note that you, as a group, a collectively responsible for the report. Your contributions should be roughly equal in amount and difficulty.

Notes

Your entire report has a **hard page limit of 4 pages** excluding references and the contribution statement. (I.e. any pages beyond page 4 must only contain the contribution statement and references). Appendices are *not* possible. But you can put additional material, like interactive visualizations or videos, on a githunb repo (use links in your pdf to refer to them). Each report has to contain **at least three plots or visualizations**, and **cite at least two references**. More details about how to prepare the report, inclucing how to produce plots, cite correctly, and how to ideally structure your github repo, will be discussed in the lecture, where a rubric for the evaluation will also be provided.

References