Mapping the American Literary Canon (1850–2006): An Exploratory Data Analysis within the Digital Humanities Framework

by Constantinos Panayi
Postgraduate Student, MSc Digital Humanities
National and Kapodistrian University of Athens, University of Cyprus, ATHENA Research Centre

***Abstract*—This study presents an exploratory data analysis (EDA) of a curated dataset of 50 canonical works of American literature spanning the mid-nineteenth to the early twenty-first century (1850–2006). Within a Digital Humanities (DH) framework, the analysis combines categorical, numerical, and textual variables to examine genre diversity, publication patterns, reader ratings, and lexical trends in work descriptions. Visual analytics and regression modelling are employed to explore relationships between textual length, publication year, and reader evaluation, while frequency-based natural language processing (NLP) techniques surface dominant thematic signals in the corpus. The findings indicate a strong plurality of genres, a concentration of works in the twentieth century, limited explanatory power of textual length for reader ratings, and a moderate contribution of publication date, while thematic vocabulary highlights narrative and socially inflected concerns across the canon. The present study was conducted as the final project for the *Data Analytics* course of the MSc Digital Humanities programme, jointly offered by the *National and Kapodistrian University of Athens*, the *University of Cyprus*, and the *ATHENA Research Centre*.**

***Index Terms*—American Literature, Literary Canon, Exploratory Data Analysis, Digital Humanities, Natural Language Processing, Linear Regression Analysis, Data Visualisation**

I. Introduction

The present study aims to develop and present a complete and coherent data analysis scenario applied to a curated dataset drawn from the American literary canon. The analysis is situated within the broader methodological framework of DH, where quantitative and qualitative approaches are jointly employed to identify trends, patterns, distributions, and relationships embedded in literary production and reception. By integrating statistical exploration, visual analytics, and text-based analysis, the study seeks to demonstrate how computational methods can contribute to the systematic examination of literary corpora while maintaining interpretative awareness of historical and cultural context.

A. Aims

Within this context, the primary objective of the project is to address six research questions focusing on:

- genre composition and diversity,
- publication trends and temporal range,
- reader evaluation and reception,
- textual length patterns,
- recurring lexical features in work descriptions.

II. Data Management

A. Dataset Overview

The American Literature Dataset consists of 50 representative and seminal works of American literature, spanning from the mid-nineteenth century to the early twenty-first century. The dataset was provided by the instructors of the *Data Analytics* course of the MSc Digital Humanities programme. This temporal breadth enables the dataset to capture a wide historical and stylistic range, encompassing major literary movements, shifts in literary form, and recurring thematic concerns across different periods. Each entry in the dataset represents a single literary work and is described by eight variables, as summarised in Table I, which collectively capture complementary dimensions of literary information. Specifically, the dataset comprises:

- four **categorical** variables related to title, authorship and literary classification;
- three **numerical** variables associated with temporal placement, textual length, and average reader evaluation; and
- one **textual** variable consisting of extended summaries and thematic discussions of each work.

The combination of these heterogeneous data types enables a multi-modal analysis and supports the application of a diverse range of data analytics methods, thereby facilitating a comprehensive examination of both quantitative patterns and qualitative characteristics within the corpus.

B. Data Quality and Pre-Processing

Before proceeding to the analysis of the selected research questions, an initial exploratory overview of the dataset is conducted in order to establish a foundational understanding of its structure, range, and internal variability [1]. As outlined above, the dataset comprises 50 observations and eight variables, encompassing a balanced combination of numerical, categorical, and textual data,

TABLE I: Dataset Variables and Descriptions

Variable Name	Description
title	The full title of the work.
author	The full name of the author of the work.
year	The year of original publication or release of the work.
pages	The length of the work measured in number of pages.
category	The specific literary genre or form of the work.
literary_period	The broader literary movement or historical period to which the work belongs.
average_rating	The average reader rating of the work on a scale from 1.0 to 5.0, representing reader reception.
description	An extended textual description summarising the plot and thematic concerns of the work.

which is well suited to multi-modal analysis. Moreover, no missing values or duplicate records are detected across any of the variables, thereby ensuring data completeness, consistency, and reliability for subsequent analyses.

Descriptive statistics and outlier detection are then applied to the numerical variables to assess their distributions and identify potential extreme values. Among the available techniques, the Interquartile Range (IQR) method is selected due to its robustness, as it relies on quartiles and is therefore less sensitive to extreme observations, making it particularly suitable for skewed distributions [1]. The results indicate that publication years are primarily concentrated between the 1920's and 1960's, reflecting a strong representation of twentieth-century American literature, while a limited number of earlier works appear as outliers that correspond to canonical rather than anomalous cases. Page counts exhibit moderate dispersion, with most works ranging between approximately 200 and 400 pages. A small number of substantially longer texts increases the mean, likely due to differences in literary form or genre. Finally, average reader ratings display relatively low variability, with values clustered within a narrow range, suggesting a generally consistent level of reader evaluation across the corpus despite minor deviations.

All detected outliers are retained in the analysis, as they correspond to historically or culturally significant works whose exclusion could compromise the interpretive validity of the results, particularly within a DH context where exceptional cases often carry analytical importance [4]. Nevertheless, certain limitations of the dataset should be acknowledged. The relatively small sample size and its curated nature may limit generalisability, while average ratings preliminarily indicate a subjective measure influenced by reader preferences. Despite these constraints, they do not diminish the dataset's value as an exploratory analytical resource, but rather frame the interpretation of the findings within an appropriate methodological and critically aware context.

III. Methodology

A. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) constitutes the primary methodological framework of the study and is employed to systematically examine the structure, distribution, and variability of the dataset prior to any modelling procedures. EDA enables the identification of dominant patterns, potential outliers, and relationships among variables through a combination of descriptive statistics and visual analytics, thereby supporting an informed and transparent analytical process.

Visualisation techniques are selected according to variable type and analytical purpose:

- **Categorical counts**: Horizontal bar plots, Pareto charts, and treemaps are used to examine the distribution and relative prominence of literary categories and genres.
- **Numerical distributions**: Box plots and histograms are employed to assess dispersion, central tendency, skewness, and the presence of outliers in numerical variables such as publication year, page count, and average rating.
- **Relationships between variables**: Scatter plots combined with fitted regression lines and residual diagnostics are utilised to explore potential associations between textual length, publication date, and reader ratings [2].

B. Modelling Approach

To investigate the extent to which measurable textual and temporal attributes contribute to reader evaluation, two linear regression models are constructed. These models are applied in an exploratory manner, with the aim of assessing explanatory trends rather than achieving high predictive accuracy [3].

1) **Simple linear regression**, modelling the relationship between the number of pages and the average reader rating, in order to examine whether textual length alone provides any meaningful explanatory power.
2) **Multiple linear regression**, incorporating both page count and publication year as predictors of average rating, allowing for the assessment of combined effects and potential temporal influence on reader reception.

Model performance is evaluated through standard regression diagnostics, including coefficient estimates, goodness-of-fit measures, and visual inspection of residuals.

C. Text Processing

Textual analysis is conducted on the descriptive summaries associated with each literary work, with the objective of identifying recurring thematic and lexical patterns across the corpus. Prior to analysis, the text data undergo a preprocessing pipeline designed to reduce noise and enhance semantic consistency. This pipeline includes lowercasing, removal of punctuation and non-alphabetic tokens, stopword elimination, and lemmatisation.

Following preprocessing, lemma frequency analysis is performed, and the forty most frequently occurring lemmas are extracted and visualised. This frequency-based approach provides an interpretable overview of dominant thematic signals within the dataset and complements the quantitative analyses conducted on structured variables.

IV. Research Questions and Findings

A. *RQ1: How many works belong to each category?*

The first research question examines how many works belong to each literary category in the dataset, aiming to assess the distribution of literary genres represented within the American literary canon. To address this question, a horizontal bar plot (Figure 1) is employed as the primary visual aid, as it constitutes an appropriate and effective visualisation technique for categorical count data [1] [2]. The horizontal orientation is deliberately chosen over a vertical bar plot due to the large number of literary categories, which would otherwise compromise label readability and overall interpretability. Categories are displayed in ascending order according to their frequency, while relative percentages are additionally reported to facilitate proportional comparison across categories. Finally, a categorical colour palette is used rather than a sequential one, emphasising that the categories represent independent and non-ordinal literary classifications rather than a continuous scale.

The results from this primary visualisation reveal a highly diverse distribution of literary categories. Specifically, 21 categories are represented by a single work, six categories by two works, four categories by three works, and one category (Southern Gothic) by five works. Although the absolute counts are relatively small, the corresponding proportions range approximately from 2% to 10%, indicating noticeable variation without extreme overrepresentation.

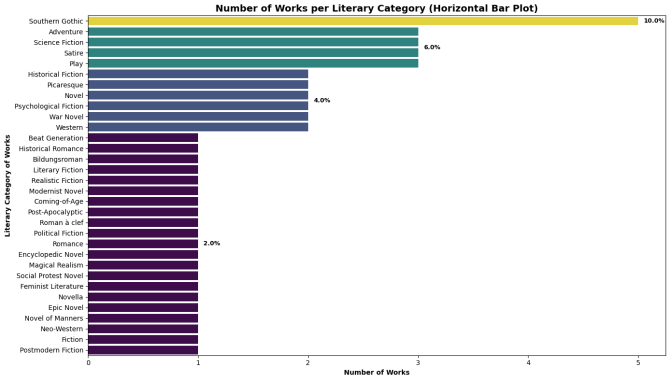

Fig. 1: Number of works per literary category (Horizontal Bar Plot).

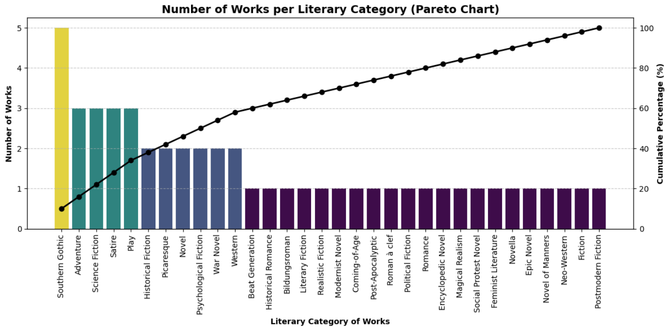

Fig. 2: Number of works per literary category (Pareto Chart).

To further support interpretation beyond simple frequency comparison, two additional visual aids are included as supplementary and intuition-oriented representations. First, a Pareto chart (Figure 2) is used to display the cumulative percentage of works per category, highlighting concentration patterns and revealing a pronounced long-tail distribution. This long tail indicates that while a small number of categories accounts for a modest share of the dataset, the majority of categories are represented by very few works, underscoring the diversity of genres included. Second, a treemap (Figure 3) provides an intuitive overview of the relative proportions of works across literary categories, reinforcing the absence of strong dominance by any single genre and visually conveying the broad variety present in the dataset. Importantly, while both the Pareto chart and the treemap enhance interpretive intuition, the bar plot remains the central analytical visual.

Overall, this distribution highlights a strong degree of genre diversity within the dataset, suggesting that American literature, as represented here, is characterised by thematic and stylistic plurality rather than dominance by a single literary form. The prominence of Southern Gothic points to the significance of regional and thematic traditions within American literary history, particularly those engaging with social tension, identity, and historical memory [5]. Other relatively frequent categories, such as Science Fiction, Satire, Adventure, and Play, further reflect genres traditionally associated with social critique, speculative imagination, and broad readership appeal. At the same time, the large number of categories represented by a single work underscores the curated nature of the dataset, emphasising the inclusion of emblematic and seminal works that function as representative examples of specific genres or movements rather than exhaustive genre coverage. However, the observed frequencies should be interpreted with caution, as they may be partly influenced by the selective composition of the dataset, rather than indicating proportional dominance within American literature as a whole.

A limitation of this analysis lies in the small number of observations per category, which restricts the potential for

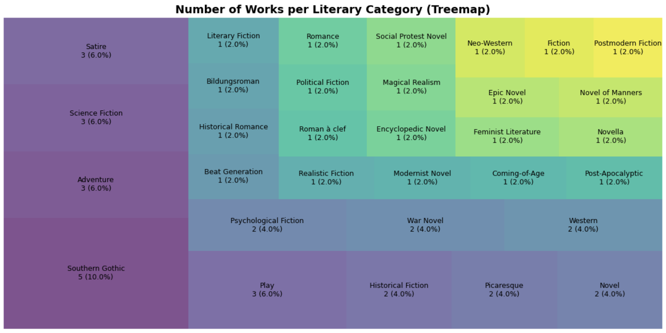

Fig. 3: Number of works per literary category (Treemap).

broad generalisation at the genre level. Additionally, while a pie chart could have been considered, the high number of categories and their relatively small proportions would significantly hinder readability and comparative clarity [2]. Consequently, the horizontal bar plot, together with the two supplementary visual aids, constitutes the most suitable set of visualisation choices for addressing this research question. Finally, deeper in-sights into the role of literary categories within the American literary canon may be achieved through their combined analysis with other variables, such as literary periods, average ratings, and textual features.

B. RQ2: What is the range of publication date of the works?

This research question investigates the temporal span of the dataset in order to identify the overall range of publication years and to detect historically early or late works that may constitute extreme cases within the dataset. To explore this question, publication years are primarily visualised using a box plot (Figure 4), which provides a concise summary of dispersion and, by extension, the dataset's range [1]. Through the IQR method, the box plot highlights the lower and upper bounds of the data as well as potential outliers beyond the conventional thresholds (Q1 - 1.5×IQR and Q3 + 1.5×IQR).

In addition, a histogram grouped by decade (Figure 5) is produced as a complementary visual aid. Rather than serving as a tool for precise range measurement, the histogram provides contextual insight into how publication years are distributed within the identified temporal bounds, thereby supporting the interpretation of concentration patterns and underrepresented periods across time. However, it should be noted that this graph does not permit an exact reading of the dataset's minimum and maximum values and is inherently influenced by the choice of bin width and grouping strategy. For this reason, it is employed supplementarily rather than as a primary instrument for range detection.

The resulting box plot indicates that the publication date range spans from 1850 to 2006, corresponding to a temporal coverage of 156 years, which constitutes the strict and direct definition of the dataset's scope. Within this span, the central 50% of works is densely concentrated between 1926 and 1967, with the median located around 1949. Three early works from the early 1850's appear as low-end outliers, extending the lower quartile (1850-1926) and slightly shifting the mean towards earlier years (approximately 1942). In contrast, the upper quartile (1967-2006) is comparatively compact, suggesting a denser representation of later twentieth- and early twenty-first-century works within the dataset.

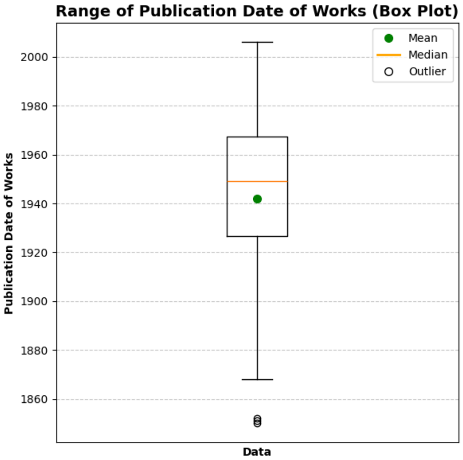

Fig. 4: Range of publication date of works (Box Plot).

In addition, the decade-based histogram corroborates the range identified by the box plot by illustrating that the majority of works fall within the 20th century, while only a small number of early publications extend the lower temporal boundary into the mid-19th century. Although limited in number, these early cases are decisive in defining the overall range and underscore the inclusion of historically foundational and canonically notable works rather than random temporal dispersion.

Overall, the visual evidence confirms that the dataset covers a publication span of more than one and a half centuries, combining a broad temporal range with a clear concentration of works from the twentieth century onwards. The dense clustering between the 1920's and the 1970's coincides with periods of intense social, economic, and political transformation, including the interwar period, the Great Depression, the Second World War, and the early Cold War era. Such contexts are often associated with heightened cultural production and strong critical reception, suggesting that literature served as a key medium for articulating collective concerns, uncertainty, and the search for meaning during times of crisis.

A more moderate yet still notable proportion of works dates from the 1980's to the early 2000's, indicating the inclusion of more recent literary trends and evolving thematic preoccupations. By contrast, the period between 1850 and 1925 is sparsely represented, pointing to a selective incorporation of early American literary history.

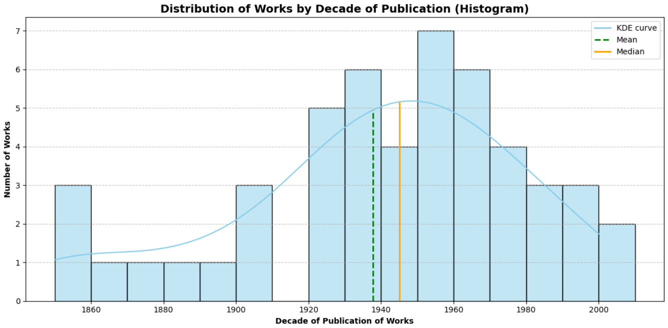

Fig. 5: Distribution of works by decade of publication (Histogram).

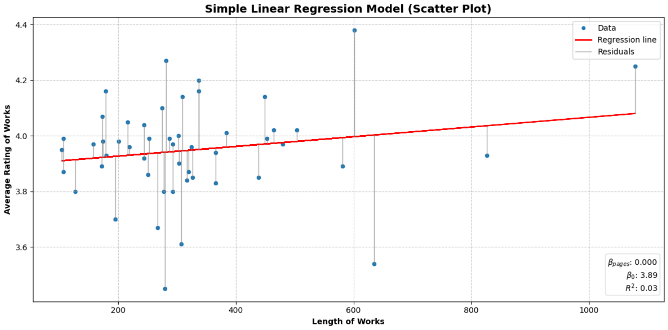

Fig. 6: Simple linear regression model (Scatter Plot).

Nevertheless, the presence of a few early canonical works indicates that historically foundational texts retain strong cultural significance despite temporal distance, functioning as emblematic anchors of the canon.

In general, analytical outcomes are inevitably shaped by methodological choices. In this analysis, outliers are deliberately retained, acknowledging that summary measures, such as the mean, are influenced by historically early works. Nevertheless, as previously discussed, excluding these cases would reduce interpretive validity within DH research, where exceptions are often analytically meaningful. Finally, future analyses could adopt more advanced approaches, such as clustering techniques (e.g., K-means), to further reveal interactions between time and content-related variation.

C. RQ3: Can we predict the rating of a work based on its length?

This research question investigates whether the average reader rating of a literary work can be predicted on the basis of its length, measured by the number of pages. Therefore, the objective is to assess whether textual extent functions as a meaningful indicator of reader reception within the dataset. To examine this question, a simple linear regression model is employed, with pages as the predictor (independent variable) and average_rating as the target (dependent variable). This relationship is then visualised using a scatter plot with a fitted regression line (Figure 6), which constitutes an appropriate visual aid for examining the direction, strength, and dispersion of a potential linear association between these two variables, while also supporting the narrative interpretation of their relationship. Vertical residual lines are also included in the plot to represent the distance between the observed data and the values predicted by the model, thereby providing a clearer visual indication of prediction error [3].

The fitted model reveals an extremely weak linear association between pages and average_rating. The slope is effectively zero, indicating that increases in the number of pages are not systematically associated with higher or lower reader ratings. The intercept (3.89) represents the baseline estimated rating around which works cluster, reflecting the mean level of reader reception irrespective of textual length. Moreover, the data points are widely dispersed and exhibit substantial deviations from the regression line, further underscoring the absence of a meaningful linear trend. This observation is reinforced by the very low coefficient of determination (0.03), which indicates that only about 3.3% of the variance in average ratings can be explained by pages alone.

Interpretively, these findings indicate that the length of a literary work does not constitute a reliable indicator of reader evaluation or reception within this dataset. Both shorter and longer works display similarly high and low ratings, implying that reader response is shaped more strongly by other factors, such as genre, thematic content, historical context, literary style, or subjective reader preferences. Several limitations should also be acknowledged. The analysis is constrained by the relatively small sample size and by the use of a single predictor within a linear frame-work. More advanced approaches, such as multiple linear regression incorporating additional variables, could offer a more nuanced understanding of the determinants of reader ratings.

D. RQ4: Which are the 40 most frequently found lemmas in the descriptions of the works?

This research question focuses on the analysis of textual data by exploring the 40 most frequently occurring lemmas in the descriptions of the literary works. The objective is to identify dominant lexical patterns and recurring thematic emphases, thereby gaining insight into the narrative and sociocultural concerns that characterise American literature as represented in the dataset.

Following standard text-cleaning and pre-processing procedures (i.e., lowercasing, noise removal, stopword elimination, and lemmatisation), lemma frequencies are computed across all work descriptions [6]. To address this question, a horizontal bar plot (Figure 7) is employed as the primary visual aid, as it constitutes an appropriate and effective visualisation technique for categorical frequency data. The horizontal orientation is deliberately chosen over a vertical bar plot due to the large number of lemmas,

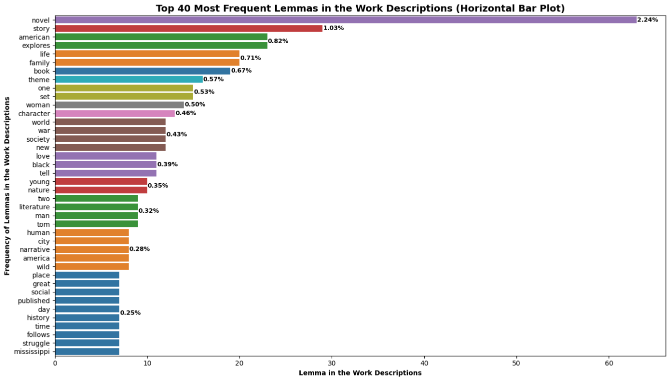

Fig. 7: Top 40 most frequent lemmas in the work descriptions (Horizontal Bar Plot).

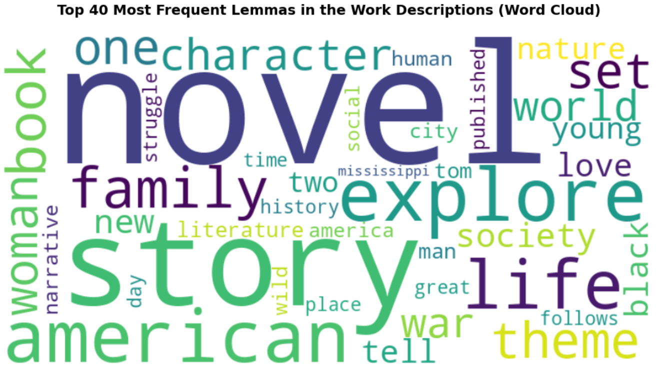

Fig. 8: Top 40 most frequent lemmas in the work descriptions (Word Cloud).

which would otherwise compromise label readability and overall interpretability. Words are displayed in ascending order according to their frequency, while relative percentages are additionally reported to facilitate proportional comparison across categories. Finally, a categorical colour palette is used rather than a sequential one, emphasising that the results represent independent and non-ordinal lexical categories rather than a continuous scale.

To further support interpretation beyond simple frequency comparison, two additional visual aids are included as supplementary and intuition-oriented representations. First, a word cloud (Figure 8) is used to provide an exploratory overview of the most salient terms, highlighting their relative prominence without additional numerical distraction [2]. Second, a treemap (Figure 9) offers an intuitive over-view of the relative proportions of the 40 most frequent lemmas, indicating that, although novel emerges as the most prominent term, no single word otherwise dominates the lexical field, which remains relatively balanced across the remaining lemmas. Importantly, while both complementary visuals enhance interpretive intuition, the bar plot remains the central analytical visual for this research question.

From an analytical perspective, the results derived from these visualisations reveal a clear dominance of narrative-related vocabulary across the corpus. The lemma novel emerges as the most frequent term, accounting for over 60 occurrences and more than 2.24% of all processed tokens, followed by story (1.03%). Other highly frequent lemmas include American, explore (0.82%), life, and family (0.71), while additional terms relating to social structures (e.g., society, social), identity groups (e.g., human, man, woman, young, black), and historical or spatial references (e.g., war, history, city, America, Mississippi). Abstract concepts (e.g., life, world, love, nature), alongside explicitly literary terms (e.g., book, narrative, literature), further underscore the thematic and cultural breadth of the descriptions.

These findings offer substantive indications of the dominant thematic and narrative axes of American literature as reflected in the descriptions. The prominence of terms such as novel, story, character, and narrative points to storytelling and fictionality as central modes of literary representation. At the same time, the frequent appearance of socially and historically charged lemmas (e.g., family, life, love, society, human, woman, and struggle) highlights a persistent preoccupation with lived experience, social relations, and both personal and collective conflict. These lexical patterns suggest that the works are commonly described (and, by extension, often interpreted) through themes of identity formation, moral tension, social and gender inequality, and the negotiation between individual agency and broader social structures.

Notably, the presence of terms such as black, war, Mississippi, and American underscores a literary tradition deeply intertwined with race, national identity, and historical trauma. References to race and place – particularly the American South, signalled by Mississippi – evoke narrative environments shaped by slavery, segregation, and regional tensions, thereby resonating with genre traditions such as Southern Gothic, already observed in a previous analysis [5]. Finally, the complementary use of a word cloud and a treemap further supports interpretation by providing an immediate sense of thematic "weight" within the corpus. Overall, the combined evidence suggests a canon marked by narrative plurality, social critique, and a sustained focus on the human subject within historically specific – and often conflictual – cultural worlds.

However, several limitations should be acknowledged. First, the selection of the top 40 terms inevitably shapes the thematic emphasis observed, privileging dominant lexical patterns while potentially marginalising less frequent but conceptually significant terms. In addition, pr-processing decisions – such as lemmatisation, the exclusion of tokens shorter than three characters, and the handling of part-of-speech distinctions – directly influence the results. The differentiation of related lemmas based on grammatical form (e.g., American vs. America, social vs. society) illustrates how alternative approaches, such as stemming, could have yielded different aggregations and

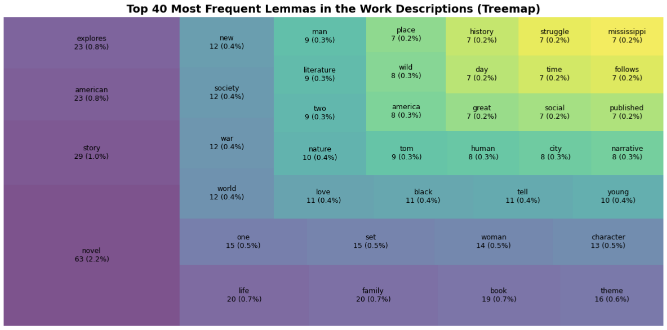

Fig. 9: Top 40 most frequent lemmas in the work descriptions (Treemap).

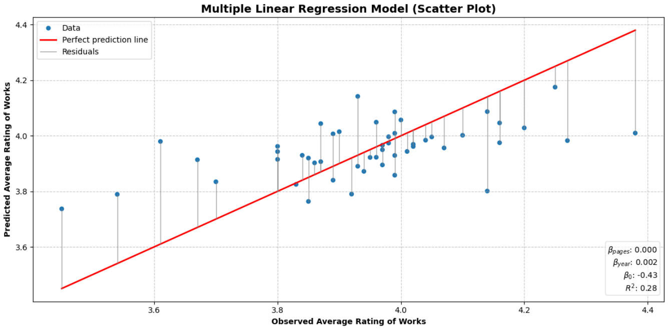

Fig. 10: Multiple linear regression model (Scatter Plot).

interpretive outcomes.

Moreover, the removal of short tokens (e.g., us, if, so), although justified by their limited semantic weight, may obscure stylistic or grammatical patterns related to narrative perspective or modality. From a visualisation standpoint, while a pie chart could have been considered, the high number of word categories and their relatively small proportions would significantly hinder readability and comparative clarity. Consequently, the horizontal bar plot, supported by the two supplementary visual aids, constitutes the most suitable set of visualisation choices for addressing this query. Finally, as frequency-based analysis captures surface-level lexical patterns rather than deeper semantic relationships, meaning that conceptually similar ideas may remain analytically fragmented, future work incorporating semantic modelling or topic analysis could provide a more nuanced understanding of the underlying thematic structures of the corpus.

E. RQ5: Can we predict the rating of a work based on its length and publication date?

This research question extends the rationale of a previous analysis by examining whether the average reader rating of a literary work can be predicted not only by its length but also by its publication date. The objective is to assess if the joint consideration of a structural feature (pages) and a temporal feature (year) improves explanatory power compared to a simple, single-predictor model.

To address this question, a multiple linear regression model is applied to predict average_rating (target variable) based on pages and year (predictors). In short, multiple linear regression extends the logic of simple regression by estimating the unique contribution of each predictor while holding the others constant, thereby modelling how different factors jointly explain variation in the outcome. This approach enables a deeper understanding of how textual length and temporal placement independently or jointly interact in shaping reader evaluation and reception.

To evaluate the predictive performance of the model, a scatter plot of observed versus predicted average reader ratings is employed as the primary diagnostic visualisation (Figure 10). This plot enables an intuitive assessment of model accuracy by displaying how closely predicted values align with observed ones relative to the perfect prediction line. Deviations from this line indicate prediction error, while the overall dispersion of points reflects the strength of the predictive relationship. Vertical residual lines are also included in the plot to visualise the magnitude and direction of individual prediction errors, thereby providing direct insight into model fit.

Complementarily, two bar plots of regression coefficients are included as supportive explanatory visualisations. The first displays unstandardised coefficients (Figure 11), which express the effect of each predictor in its original units and thus facilitate substantive interpretation. The second presents standardised coefficients (Figure 12), which place all predictors on a common scale, enabling direct comparison of their relative importance within the model. Together, these plots clarify not only whether the model has predictive power, but also which predictor contributes more strongly to that power [3].

Overall, the fitted model indicates a moderately weak linear relationship be-tween the predictors and the target variable. The coefficient for pages (0.0001) is effectively negligible, indicating that increases in textual length have no meaningful impact on average reader ratings. By contrast, the positive coefficient for year (0.002) suggests that more recently published works tend to receive slightly higher ratings. According to the coefficient of determination, approximately 28.17% of the total variance in average_rating is explained jointly by year and pages. Although this level of explanatory power remains limited, it constitutes a substantial improvement over the previous simple linear regression, corroborating that the inclusion of a second predictor enhances model performance.

The relative impact of each predictor is further clarified by the coefficient bar plots. Both the unstandardised and standardised representations show that year exerts a noticeably stronger influence on predicted ratings than pages, confirming the pattern observed in the regression output. Moreover, the observed-versus-predicted scatter plot further reveals that the vast majority of predicted ratings cluster between 3.8 and 4.1, even when observed

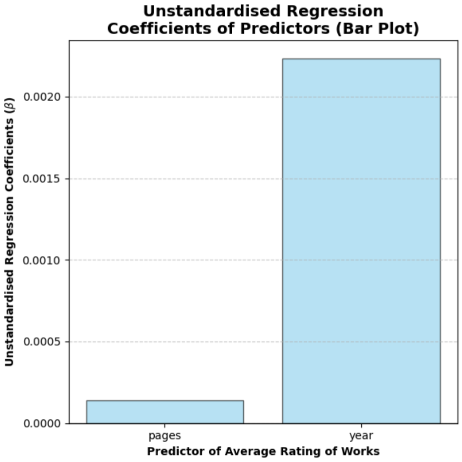

Fig. 11: Unstandardised regression coefficients of predictors (Bar Plot).

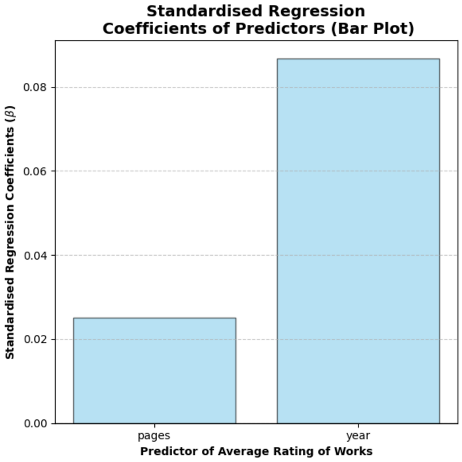

Fig. 12: Standardised regression coefficients of predictors (Bar Plot).

values extend beyond this range. Lower observed ratings tend to be overestimated, while higher observed ratings tend to be underestimated, indicating a regression-to-the-mean effect. This trend suggests that, within this model, reader ratings are pulled toward a relatively narrow central band, reflecting the limited sensitivity of ratings to variation in length and publication date alone.

Taken together, these findings suggest that publication date contributes more strongly to reader evaluation than textual length, potentially reflecting contemporary reader preferences, greater cultural familiarity, or a bias toward more recent works in rating practices. Consistent with earlier results, length alone does not function as a reliable predictor of reader reception. Although the inclusion of a second predictor improves explanatory power, the majority of variation in average_rating remains unexplained, indicating that reader evaluation is shaped primarily by qualitative, thematic, and subjective factors, such as narrative content, historical and cultural significance, and individual taste.

However, several limitations should be acknowledged. The relatively small sample size constrains statistical power and limits generalisability. Average ratings represent also a subjective measure influenced by reader bias and contemporary cultural norms. Moreover, the model excludes potentially influential variables, such as genre, literary period, or thematic complexity, which may exert stronger effects on reader reception. Within a DH context, the results should therefore be understood as exploratory rather than predictive, underscoring the need for future research employing multi-factor and semantically informed models.

F. RQ6: How do the ratings of the works vary across literary periods?

This research question introduces an additional analytical dimension to the study of reader reception by examining how average_rating vary across different literary periods. The aim is to investigate whether specific literary movements are associated with systematically higher or lower reader evaluations, thereby exploring the potential role of literary period as a contextual factor influencing reception.

To address this research question, a complementary set of three visualisations is employed: a box plot, a swarm plot, and a violin plot. Each visualisation serves a distinct analytical purpose, while jointly mitigating the limitations of the others. First, a box plot (Figure 13) is used as the primary analytical visualisation to summarise the central tendency, range and potential outliers of average_rating across literary periods. Through the IQR method, the box plot provides a concise and standardised comparison of variability and median values between periods, allowing for an initial assessment of systematic differences in reader ratings. However, despite its strengths, the box plot is not fully sufficient on its own, as it conceals the underlying number of observations and may therefore be misleading in categories with small sample sizes. In periods represented by only one or two works, box plot statistics can appear artificially stable or comparable, despite lacking statistical robustness [1].

To address this limitation, a swarm plot (Figure 14) is employed as a supportive visualisation. By displaying individual data points for each literary period, the swarm plot enables direct inspection of sample size, clustering, spread, and extreme values. This representation is particularly important for identifying underrepresented periods, where limited observations prevent strong generalisations. In such cases, this plot explicitly reveals the fragility of any apparent patterns, thereby reducing the risk of overinterpretation derived from summary statistics alone.

Finally, a violin plot (Figure 15) is used as a more intuitive and exploratory visualisation to examine the distributional shape and density of average ratings within each literary period. By combining it with the features of box plot, the violin plot offers further insight into symmetry, concentration, and multimodality, supporting a more informative reading of how ratings are distributed across periods. Although alternative visualisations, such as bar plot, could have been employed, they are not selected, as they are less informative when comparing multiple groups of unequal and limited size. Apart from this, mean-based summaries can be misleading in such contexts, as they fail to capture distributional shape and may overrepresent categories with very few observations [2].

Overall, the combined visualisations reveal noticeable variation in average ratings across literary periods. Contemporary and Postmodernist works exhibit comparatively higher ratings, with values largely ranging between 3.9 and 4.2. Contemporary literature, in particular, shows a high central tendency and relatively limited dispersion, albeit with the presence of an extreme high value, while Post-modernist works display a broader spread of ratings. Modernism, the most frequently represented movement in the dataset, demonstrates the widest range of values, spanning approximately from 3.7 to 4.4 and including several high-end observations. Realism presents a moderately dispersed distribution, with ratings densely concentrated around the 3.9-4.0 range, whereas Post-War literature tends to exhibit slightly lower average ratings overall (approximately 3.6-4.0). Finally, periods such as Romanticism, American Renaissance, Harlem Renaissance, and Naturalism appear in the box and violin plots, respectively, but, as clearly revealed by the swarm plot, are represented by only one or two works each. As a result, any apparent distributional patterns for these periods should be interpreted with caution, as they are statistically unstable and analytically fragile due to underrepresentation.

Interpretively, the results suggest that more recent literary periods, particularly Contemporary and Postmodernism, tend to receive slightly higher and more consistently positive reader ratings. This pattern might reflect temporal proximity, greater cultural familiarity, or alignment with contemporary reader expectations and sensibilities. At the same time, the substantial dispersion observed within Modernism and the strong performance of Realist works indicate that literary period alone does not determine reader reception. Older movements can still achieve high ratings, suggesting that thematic relevance, narrative quality, and canonical status remain influential irrespective of historical distance. Overall, the findings point to a complex and non-deterministic relationship between literary period and reader evaluation, where period may interact with other factors rather than acting as a primary driver of ratings.

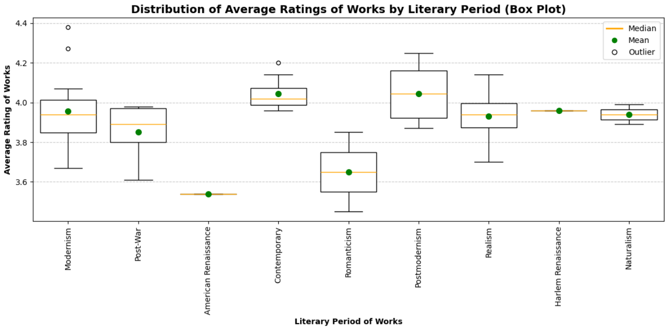

Fig. 13: Distribution of average ratings of works by literary period (Box Plot).

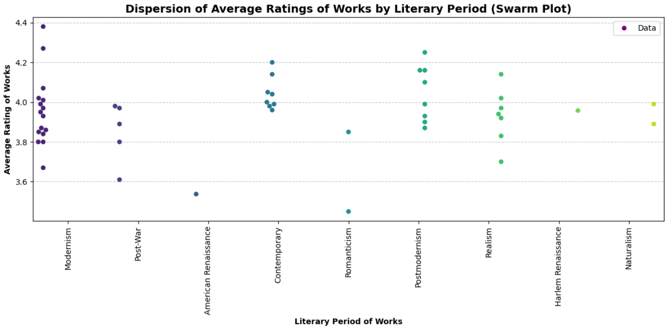

Fig. 14: Dispersion of average ratings of works by literary period (Swarm Plot).

Therefore, several limitations must be acknowledged. First, the small and uneven sample sizes across periods substantially restrict generalisability and increase the risk of over-interpretation, especially for underrepresented categories (e.g., Romanticism). Second, reader ratings are inherently subjective and influenced by unobserved variables such as reader age, cultural background, or familiarity with specific literary movements. Third, the curated nature of the dataset limits statistical inference and favours exploratory rather than confirmatory conclusions. Within a DH framework, the results should therefore be interpreted as contextual and indicative, highlighting patterns worthy of further investigation rather than definitive evaluative hierarchies across literary periods.

V. Conclusion

The present study set out to develop and demonstrate a complete data analysis workflow applied to a curated dataset of the American literary canon, integrating numerical, categorical, and textual data within a DH framework. Through a sequence of exploratory analyses, visualisations, and statistical models, the study addressed six research questions focusing on distributions, relationships, and thematic patterns across literary works.

Overall, the findings highlight the value of quantitative and computational methods as interpretive and question-generating tools, rather than deterministic explanatory mechanisms, in literary research. The analysis of publication dates reveals a strong concentration of works in the twentieth century – particularly between the 1920's and the 1970's – reflecting historically and culturally formative periods in American literature. Genre- and period-based

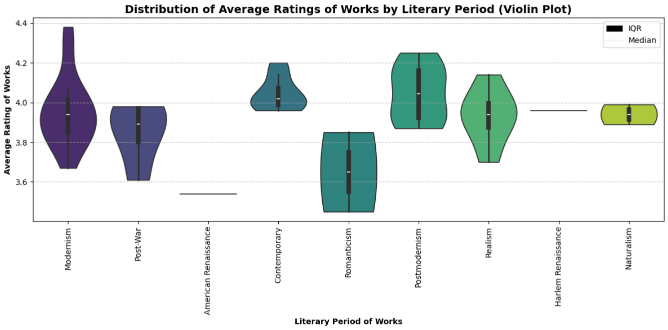

Fig. 15: Distribution of average ratings of works by literary period (Violin Plot).

analyses further demonstrate a high degree of thematic and stylistic plurality, underscoring that American literature cannot be reduced to a single dominant form, but instead consists of overlapping movements, regional traditions, and narrative concerns.

The examination of reader ratings indicates that simple quantitative features, such as textual length, have minimal explanatory power when considered in isolation. Although publication date and literary period introduce moderate variation – favouring more recent movements – reader reception remains primarily shaped by qualitative, contextual, and subjective factors rather than measurable attributes alone. Moreover, textual analysis of work descriptions further complements these insights. The prominence of narrative-related lemmas, alongside socially and historically charged terms, highlights the centrality of storytelling, identity, and social struggle within the American literary canon. Importantly, this analysis demonstrates how even relatively simple NLP techniques can surface dominant thematic axes, while simultaneously revealing their limitations in capturing deeper semantic structures.

Several methodological limitations must therefore be taken into consideration. The relatively small and curated dataset restricts generalisability and amplifies the risk of over-interpretation, particularly for underrepresented features. Visualisations such as box plots and word clouds, while effective for exploratory analysis, may obscure sample-size imbalances or exaggerate apparent differences. In addition, pre-processing decisions in textual analysis – such as lemmatisation strategies, frequency thresholds, or the exclusion of short tokens – inevitably shape analytical outcomes.

Despite these constraints, the study demonstrates the epistemic value of data analytics in DH as an exploratory and reflexive practice. Rather than replacing close reading or philological scholarship, computational methods function as lenses that reveal patterns, prompt new queries, and support comparative reasoning. In this sense, the analytical process itself – along with its assumptions, choices, and limitations – becomes an integral part of the interpretive outcome. In conclusion, this work demonstrates that data-driven approaches, when applied critically and transparently, can meaningfully enrich literary analysis by bridging quantitative patterns with qualitative interpretation, thereby reinforcing the interdisciplinary potential of DH research.

Acknowledgements

The author would like to thank the instructors of the *Data Analytics* course for providing the dataset and methodological guidance. The author also gratefully acknowledges the *Union of Greek Shipowners* and the *Georgios and Andrianí Lordos Foundation* for the financial support provided through the award of scholarships that enabled the completion of his postgraduate studies.

References

[1] Canning, J. (2014) *Statistics for the Humanities.* John Canning.

[2] The Data Visualisation Catalogue (no date). Available at: https://datavizcatalogue.com. Accessed: 30 January 2026.

[3] Ober, P. B. (2013) “Introduction to linear regression analysis”. *Journal of Applied Statistics*, 40(12), pp. 2775–2776. Available at: https://doi.org/10.1080/02664763.2013.816069. Accessed: 30 January 2026.

[4] Schöch, C. (2013) “Big? Smart? Clean? Messy? Data in the Humanities”. *Journal of Digital Humanities*, 2(3). Available at: https://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/. Accessed: 30 January 2026.

[5] Wikipedia (2026) *American literature.* Available at: https://en.wikipedia.org/wiki/American_literature. Accessed: 30 January 2026.

[6] Piotrowski, M. (2012) “Natural Language Processing for Historical Texts”. *Synthesis Lectures on Human Language Technologies*, 5. Available at: https://link.springer.com/book/10.1007/978-3-031-02146-6. Accessed: 30 January 2026.