CS 210 DATASCIENCE PROJECT – GOODREADS DATA ANALYSIS

Defne Kızılkaya 31279

MOTIVATION

As an avid reader, l have always been intrigued by the subtle yet significant ways the changing seasons affect my reading habits. This curiosity led me to initiate a data-driven project aimed at uncovering the patterns and nuances of-my literary journey through different times of the year. I believe that seasonal changes, with their distinct emotional and environmental characteristics, play a pivotal role in influencing not only the volume of my reading but also the choice of authors, and book lengths.

Winter's cozy ambiance might draw me towards more introspective novels, while summer's vibrancy could shift my preference to lighter, more adventurous reads. This project is not just an exploration of my personal reading trends; it's a deeper dive into understanding how external factors like weather, holidays, and daylight hours intertwine with my psychological inclinations to shape my literary choices. By analyzing data from my reading history, I aim to gain insights into these seasonal reading patterns, providing a fascinating reflection of how external environments can subtly influence our personal preferences and behaviors.

HYPOTHESIS: Seasonal changes significantly influence my reading habits, with a measurable difference in the quantity of books read.

DATA ANALYSIS: TECHNIQUES USED IN DIFFERENT STAGES OF ANALYSIS

DATA SOURCE/ DATA COLLECTION

I extracted data from Goodreads. I have been using Goodreads since 2016 and I save the books that I read. Exported my reading habits as a csv file from Goodreads library.

EXPLORATORY DATA ANALYSIS

First, I imported the pandas library for data manipulation and loaded the CSV file into a pandas data frame. Then cleaned the data by dropping the unused columns.

Displayed the first few rows of the data frame with data.head() function. Used the info() function to get the summary of the data. Then described the data to generate descriptive statistics that summarizes the shape of the dataset's numerical features such as count, mean, standard deviation and, min.

Author Analysis: The dataset was grouped by 'Author' to count unique titles and identify the most read author and most read authors per season.

Yearly Reading Habits: The 'Date Read' column was converted to datetime format, and the year was extracted to analyze the number of books read each year.

Monthly Reading Trends: Similarly, the month was extracted from 'Date Read' to count the number of books read in each month, understanding monthly reading patterns.

Seasonal Analysis: The data was further manipulated to map each month to a corresponding season (Winter, Spring, Summer, Fall). The analysis involved grouping the data by year and season and counting the number of books read in each season. Also the average pages per book in each season and the average of book ratings distribution on seasons in analyzed.

Statistical Testing: An ANOVA test was performed to statistically examine if there were significant differences in the number of books read across different seasons. This involved preparing separate datasets for each season and conducting the one-way ANOVA test, provided there was sufficient data for each season.

VISUALIZATION

LINE CHARTS

Converted the ‘Date Read’ column to datetime by extracting the year, counted the occurrence of each year in the year read column which represents the

number of books read each year then sorted these counts based on the year index in ascending order. Created a line chart using Matplotlib to display the number of books read each year and each month.

Converted the ‘Date Read’ column to datetime by extracting the month and year from each date I read a book for determining when I read the most. With a mapping from months to seasons I grouped the data by ‘Season’ to count the number of books read per season. This allowed me to categorize my reading habits by the seasonal context. I used stacked bar chart to visually represent the data. This gave me an immediate visual insight into how my reading habits have fluctuated over the years with the changing seasons. This visualization helps me understand if I tend to read more in certain seasons and could potentially inform me about the influence of seasonal changes on my reading preferences.

BAR CHARTS

I used bar chart to identify and visualize the top 5 most read authors, providing a visual representation of which authors' works are most prevalent in the reading dataset. This analysis is useful for understanding reading preferences in terms of authors, highlighting which authors are most popular or most frequently read in the dataset.

Created a bar chart using the data from seasonal_counts. This visualization helps in understanding how the number of books read varies across different seasons in general.

STACKED BAR CHARTS

I used a stacked bar chart visualizing the number of books read by me per season for each year, using different colors for seasons and including a legend.

Generated a stacked bar chart that visualizes the number of books read in each season over different years from the seasonal_counts data. Each bar represents a year, with segments stacked to show the count of books read in each season. The 'viridis' colormap assigns different colors to different seasons. This visualization helps in understanding seasonal reading patterns over the years, showing how reading habits might change with seasons.

BOX PLOT

A boxplot is created using Seaborn to display the distribution of rating ('My Rating') for each season. This visualization helps identify patterns, outliers and the spread of my ratings of books in different seasons.

MACHINE LEARNING

I used machine learning for detailed analysis and prediction of the number of books read per season for each year using linear regression models. The models are used to predict the number of books that will be read in each season for a future year (e.g., 2025).These predictions are rounded to the nearest integer and displayed. The findings of the code can be summarized as follows:

Data Preparation and Seasonal Grouping:

- The 'Date Read' column in the data is used to extract the year and month, and then the months are mapped to corresponding seasons.

- The data is grouped by both year and season to count the number of books read, creating a structured dataset for analysis.

Modelling and Prediction:

- For each season, a linear regression model is built to predict the number of books read based on the year.

- The dataset is split into training and testing sets for model validation.

- The model's performance for each season is evaluated using Mean Squared Error (MSE). A lower MSE indicates a more accurate model.

Model Performance Analysis:

- The code concludes with an analysis of each model's MSE.

- A threshold for MSE is set (in this case, 10) to determine whether a model's performance is acceptable.

- Models with MSE below the threshold are considered to be performing well, while those with higher MSE need improvement.

Creating this model provided insights into my seasonal reading trends over the years and how they may evolve in the future. It also offered an evaluation of the reliability of these predictions based on the Mean Squared Error of each model.

I used one way ANOVA (Analysis of Variance) test to perform a statistical analysis to determine whether there are significant differences in the number of books read across different seasons.

Here's a breakdown:

1. **Data Preparation**:

- I converted the 'Date Read' column in the DataFrame data to datetime format, with any errors coerced into NaT (Not a Time). Extracted year and month from the 'Date Read' column. Mapped months to their respective seasons using a predefined dictionary.

2. **Grouping and Counting**:

- Then grouped the data by both year and season, and counted the size (count) of each group, representing the number of books read each season of each year.

3. **Preparation for ANOVA Test**:

- Seperated the book counts into different series based on the season (Winter, Spring, Summer, Fall).
- Performed a check to ensure that there are sufficient data points (more than one) for each season, which is a prerequisite for conducting an ANOVA test. If there's sufficient data, a one-way ANOVA test is conducted to compare the means of the number of books read across different seasons.

4. **One-way ANOVA Test**

- Via test calculated an F-value and a P-value.
- The F-value indicates the ratio of variance between the groups (seasons) to the variance within the groups.
- The P-value indicates the probability of observing the data if the null hypothesis (no difference in means across seasons) is true.

5. **Results Interpretation**:

The results from the ANOVA test, with an F-Value of approximately 5.54 and a P-Value of about 0.0055, provided valuable insights:

F-Value: The F-Value in ANOVA indicates the ratio of the variance between the groups (different seasons in my case) to the variance within the groups. A higher F-Value typically suggests a larger variation between the groups compared to within them. So an F-Value of 5.54 suggests that there is a notable difference in the means of the groups.

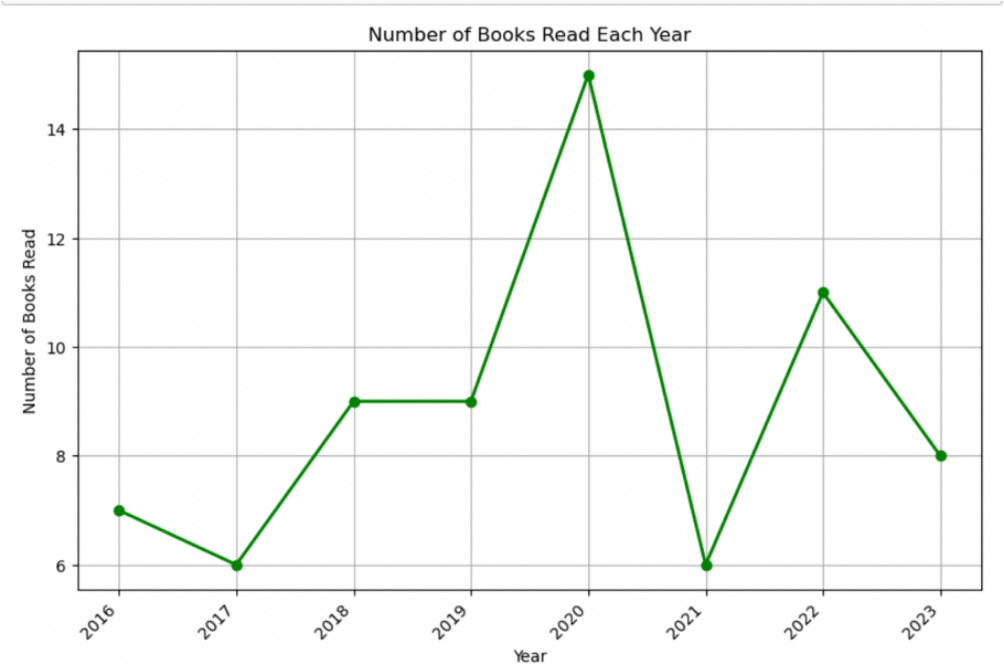
P-Value: The P-Value helps determine the statistical significance of the results. A P-Value of 0.0055 is less than the conventional threshold of 0.05 significance level. This indicates that the results are statistically significant. Since the P-Value is less than 0.5, it shows that there is a statistically significant difference in the number of books read across different seasons. This finding supports my hypothesis that " Seasonal changes significantly influence my reading habits, with a measurable difference in the quantity of books read."

FINDINGS

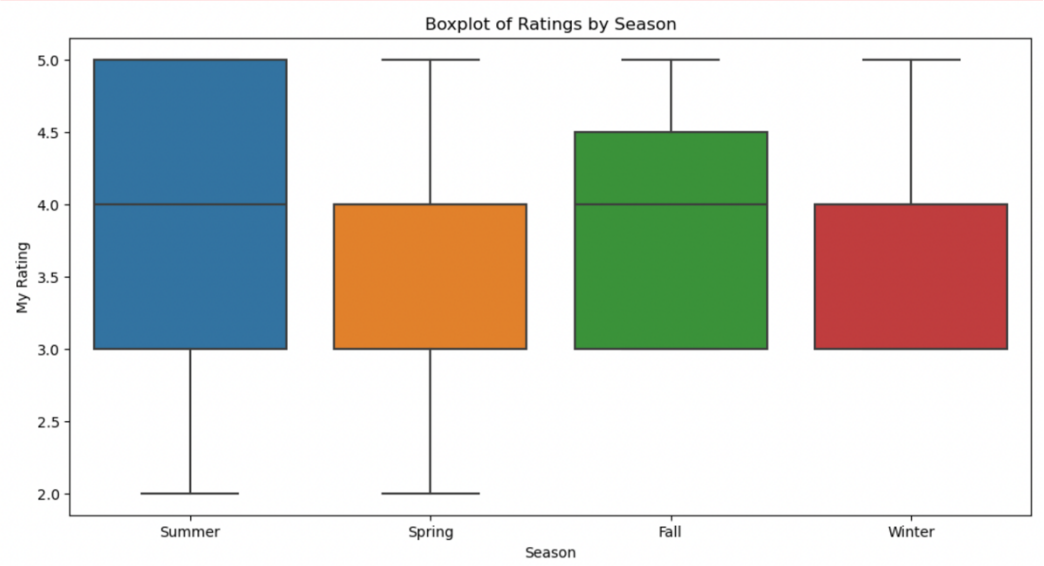
Through this project, I embarked on a methodical exploration of my reading history, and the data has revealed several enlightening trends that correlate with the rhythm of the seasons. I learned that my reading volume and preferences are indeed influenced by seasonal changes, confirming my initial hypothesis.

My venture into machine learning through linear regression models provided predictive power to my project, offering a quantitative glimpse into future reading trajectories based on past patterns. Surprisingly, the models suggested a steady, reliable reading pace that might resist the test of seasonal shifts.
However, the most striking revelation came from the ANOVA test results, which emphasized a statistically significant variance in the volume of reading across different seasons. This statistical affirmation of seasonal impact went beyond mere numbers; it was a quantifiable validation of the subtle interplay between the environment and my personal behavior.

In essence, this project has been a mirror to my inner reader, revealing how much my surroundings sway my literary appetites. It has been an exercise in self-awareness, demonstrating the power of data to illuminate the nuances of personal preference and habit.



I was curious about the yearly analysis of my reading habit as well and when I saw the visualization of number of books read each year, I saw a significant boost in the number of the books I read in year 2020 which was the pandemic year suggesting a correlation between increased reading activity and the extended periods of isolation experienced during that time. Conversely, 2021 demonstrated a significant decline in the number of books read. This downtrend can be attributed to the rigorous academic preparation for the YKS examination which reallocates of time away from leisure reading. These observations affirm that significant external events exert a considerable influence on my reading volume.



ANOVA test for comparing mean ratings across seasons: F-Statistic=0.233332880869371 26, p-value=0.8728446296095647

I wanted to test whether the ratings I give to books follow a particular trend and do I tend to give higher rating in a specific season. I conducted an ANOVA test to find out. The results of ANOVA test with an F-Statistic of 0.2333 and a p-value of 0.8728, showed me that statistically, there is no evidence to suggest that the mean ratings for books vary significantly across different seasons. In other words, the season in which a book is read does not appear to have a significant impact on its rating.

```
Predicted books for each season in 2025 (rounded to nearest integer):
Fall: 4
Spring: 3
Summer: 4
Winter: 1
MSE for each season:
Fall: 0.6560000000000072
The model for Fall is performing well with a low MSE of 0.6560000000000072.
Spring: 7.754000000000005
The model for Spring is performing well with a low MSE of 7.754000000000005.
Summer: 3.874000000000003
The model for Summer is performing well with a low MSE of 3.874000000000003.
Winter: 1.8499999999999603
The model for Winter is performing well with a low MSE of 1.8499999999999603.
```

```
F-Value: 5.543321889227768, P-Value: 0.005460981751195989
```

The predictive analysis for the year 2025 provides a projection of my seasonal reading patterns. This projection suggests a marked preference or availability for reading during the Fall and Summer months, a moderate engagement in Spring, and a notable reduction in Winter.

The Mean Squared Error (MSE) for each seasonal model offers insight into the performance and reliability of these predictions. Each model indicates a low error rate and based on these results, it appears that the models are performing well, suggesting confidence in the robustness of the predictive outcomes.

The findings underscore a significant seasonal influence on my reading habits. Which confirms my hypothesis: Seasonal changes significantly influence my reading habits, with a measurable difference in the quantity of books read. While

external factors and personal circumstances certainly contribute to these patterns, the data-driven models offer a quantifiable validation of my intrinsic seasonal preferences.

The lower reading count in Winter could be attributed to the busy end-of-year activities or a reflection of a time when I engage in other forms of leisure. Moving forward, this analysis can be instrumental in planning my reading schedule, ensuring a consistent reading habit throughout the year, and perhaps encouraging a more proactive selection of reading material for the Winter season to balance the yearly distribution. It also invites a more profound reflection on how I allocate time for reading and how I can adapt my habits to foster continual personal and intellectual growth across all seasons.

LIMITATIONS AND FUTURE WORK

Reflecting on the current scope of my project, I recognize that the dataset utilized could be broadened to enhance the depth and reliability of the analysis. The inclusion of additional features, such as book genres, would likely yield richer insights into my reading preferences and their seasonal fluctuations.

For future endeavors, I am committed to expanding the dataset by consistently logging more detailed aspects of my reading experiences, including genres, author styles, and thematic elements. This expanded dataset will allow for a multifaceted

analysis, potentially uncovering correlations between specific genres or authors and seasonal moods or activities.

Continuing to grow the dataset will also enable the application of more sophisticated machine learning models. These models could reveal complex patterns and trends that a larger and more feature-rich dataset supports. Over time, I aspire to refine the predictive capabilities of the models further and explore the nuanced dynamics of my reading habits in even greater detail.