#### INT232:DATA SCIENCE TOOLBOX :

#### R PROGRAMMING

#### PROJECT REPORT

(Project Semester April 2024)

***ANALYSIS DASHBOARD***

***For AmazonBooks and Billboard***

Submitted by Akash Kumar 12106958

Computer Science and Engineering

#### INT232

Under the Guidance of Assistant Professor Ramanpreet Kaur :28846

**Discipline of CSE/IT**

**Lovely School of Computer Science Lovely Professional University, Phagwara**



#### CERTIFICATE

This is to certify that Akash Kumar bearing Registration no. 12106958 has completed **INTB233** project titled, **“*ANALYSIS DASHBOARD*”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor Designation of the Supervisor**

**School of …………………………………………….**

Lovely Professional University Phagwara, Punjab.

Date: 26-04-2024

Acknowledgement:

We would like to express our sincere gratitude to Ramanpreet Kaur for their guidance and support. Their teachings was instrumental in making the dashboard.

Special thanks go to School of Computer Science and Engineering for introducing the course. Their quality brought specific outcome to fruition.

Lastly, we acknowledge the support and encouragement of the teachers for their commitment to upskills in particular domian. Their guidance has been invaluable throughout this project.

This project would not have been possible without the collective efforts of everyone involved. We are truly grateful for your contributions and dedication.

Sincerely,

Akash Kumar 12106958

Table of Content

1. **Introduction**
2. **Objectives/Scope of the Analysis**
3. **Source of dataset**
4. **ETL process**
5. **Analysis on dataset (for each analysis)**
   1. **Introduction**
   2. **General Description**
   3. **Specific Requirements, functions and formulas**
   4. **Analysis results**
   5. **Visualization**
6. **List of Analysis with results**
7. **Bibliography**

# Introduction

# Welcome to the interactive dashboard for analyzing Amazon Books and Billboard datasets! In this dashboard, we delve into the fascinating world of literature and music, exploring the trends, patterns, and insights hidden within these rich datasets

# Libery Used:

# library(flexdashboard)

# library(tidyverse)

# library(highcharter)

# library(gt)

# library(htmltools)

# library(viridis)

# Objectives/Scope of the Analysis

1. Who are the most popular authors from 2009 to 2019 by there people number of reviews from amazonbooks dataset?
2. what are the name and authors of the best books from 2009 to 2019 from amazonbooks dataset?
3. which is the most comman Genre(fiction /non-fictional) from 2009 to 2019 from amazonbooks dataset?
4. which is the most popular songs by number of weeks on board from Billboard dataset?
5. which is the most popular Artists having more number of weeks on board from Billboard dataset?
6. Building the Dashboard

# Source of dataset:

The source of the dataset is picked from Kaggle.com .Its data of the sales of Amazon Books and the Billboard songs.

# ETL process:

The ETL process stands for Extract, Transform, Load. It's a fundamental component of data warehousing and analytics, involving three key stages:

**1. Extract**

Data Sources:

**Amazon Books Dataset**: Retrieve the latest version of the Amazon Books dataset from the designated source. This dataset typically comes in structured formats such as CSV or Excel files and contains attributes such as book titles, authors, ratings, reviews, and genres.

**Billboard Dataset**: Obtain the Billboard dataset from a reliable source. This dataset is often available in CSV or similar formats and includes information on top songs, chart rankings, artists, genres, and release dates.

Extraction Method:

Utilize R programming libraries such as readr or readxl to import the datasets into your R environment.

Ensure data integrity by verifying column names, data types, and any potential inconsistencies.

**2. Load**

**Data Storage:**

Load the extracted datasets into your R environment for further processing and analysis.

Consider storing the datasets in appropriate data structures such as data frames for efficient manipulation and exploration.

Loading Method:

Use functions like read\_csv() or read\_excel() to read the datasets into R data frames.

Check for any missing or null values and handle them appropriately, either by imputation or removal, to maintain data quality.

**3. Transform**

**Data Preparation:**

Preprocess the datasets to make them suitable for analysis and visualization.

Perform data cleaning tasks such as removing duplicates, standardizing categorical values, and handling outliers.

**Transformation Techniques:**

Utilize R packages like dplyr and tidyr for data manipulation tasks such as filtering, sorting, and summarizing.

Create derived variables or calculated fields to extract additional insights from the data, such as aggregating book sales by genre or calculating song popularity scores.

**Data Integration:**

Combine the Amazon Books and Billboard datasets, if applicable, to perform comparative analysis and uncover relationships between literature and music.

Merge the datasets based on common attributes such as genre or publication date to facilitate cross-referencing and correlation analysis.

**Data Validation:**

Validate the transformed datasets to ensure accuracy and consistency.

Compare summary statistics, distribution plots, and cross-tabulations against expectations and domain knowledge to verify the integrity of the transformed data.

# Analysis on dataset (for each analysis):

**1. Most Popular Authors from 2009 to 2019 by Number of Reviews**

**i. Introduction:** The objective of this analysis is to identify the most popular authors based on the number of reviews received for their books from 2009 to 2019.

**ii. General Description:** We will delve into the Amazon Books dataset to determine which authors garnered the highest number of reviews over the specified time period.

**iii. Specific Requirements, Functions, and Formulas:**

Filter the Amazon Books dataset for books published between 2009 and 2019.

Group the data by author and calculate the total number of reviews for each author.

Sort the authors based on their review counts in descending order.

**iv. Analysis Results:**

The authors with the highest number of reviews will be identified, showcasing the most popular authors from 2009 to 2019 based on reader feedback.

**2. Name and Authors of the Best Books from 2009 to 2019 from Amazon Books Dataset**

i. **Introduction:** This analysis aims to determine the best books published between 2009 and 2019 based on various factors such as ratings, reviews, and sales rankings.

ii**. General Description**: We will utilize the Amazon Books dataset to identify top-performing books in terms of ratings, reviews, and sales ranks.

iii**. Specific Requirements, Functions, and Formulas:**

Filter the dataset for books published between 2009 and 2019.

Consider multiple factors such as average ratings, review counts, and sales rankings to determine the best books.

Apply appropriate ranking algorithms or scoring formulas to prioritize book selection.

**iv. Analysis Results:**

The names and authors of the best books will be revealed, highlighting standout literary works from the specified time period.

**3. Most Common Genre (Fiction/Non-fiction) from 2009 to 2019 from Amazon Books Dataset**

i. **Introduction:** This analysis aims to identify the most common genre, whether fiction or non-fiction, among books published between 2009 and 2019.

ii. **General Description:** We will analyze the genre distribution within the Amazon Books dataset to determine the predominant category over the specified timeframe.

iii. **Specific Requirements, Functions, and Formulas:**

Extract genre information from the dataset.

Count the occurrences of each genre and determine the most frequent category.

iv. **Analysis Results:**

The most common genre, whether fiction or non-fiction, will be identified, providing insights into reading preferences during the specified period.

**4. Most Popular Songs by Number of Weeks on Board from Billboard Dataset**

i. **Introduction:** In this analysis, we aim to identify the most popular songs based on the number of weeks spent on the Billboard charts.

ii. **General Description:** We will analyze the Billboard dataset to identify songs that have achieved the longest duration on the charts.

iii**. Specific Requirements, Functions, and Formulas:**

Extract song titles and corresponding weeks on the charts from the dataset.

Rank songs based on the total number of weeks on the charts.

iv. **Analysis Results:**

The most popular songs, as determined by the number of weeks on the Billboard charts, will be revealed.

**5. Most Popular Artists with the Most Number of Weeks on Board from Billboard Dataset**

i**. Introduction:** This analysis aims to identify the most popular artists based on the total number of weeks their songs have spent on the Billboard charts.

ii. **General Description:** We will analyze the Billboard dataset to determine which artists have achieved the highest cumulative number of weeks on the charts.

i**ii. Specific Requirements, Functions, and Formulas:**

Extract artist names and corresponding weeks on the charts from the dataset.

Aggregate the weeks on the charts for each artist.

Rank artists based on the total number of weeks on the charts.

**iv. Analysis Results:**

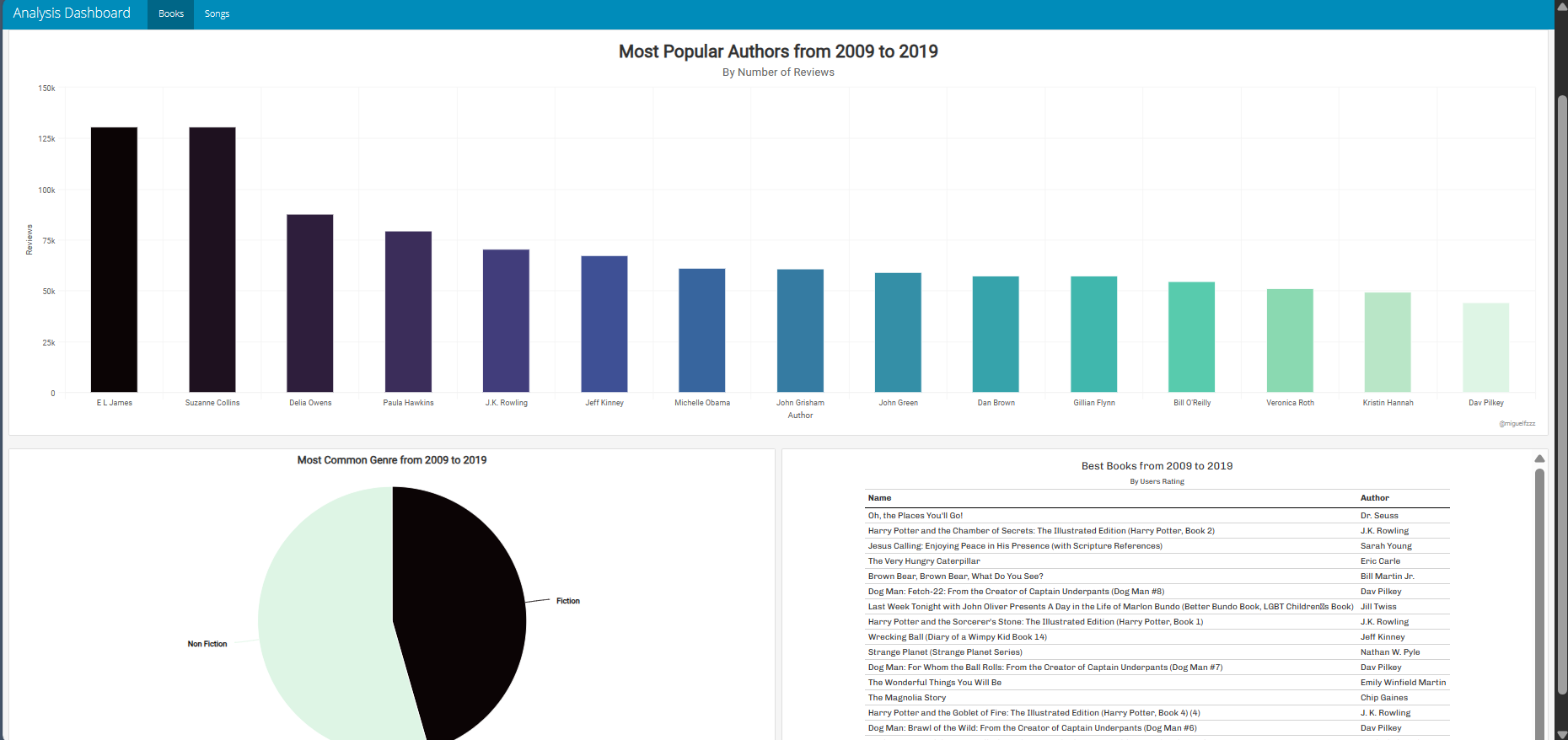
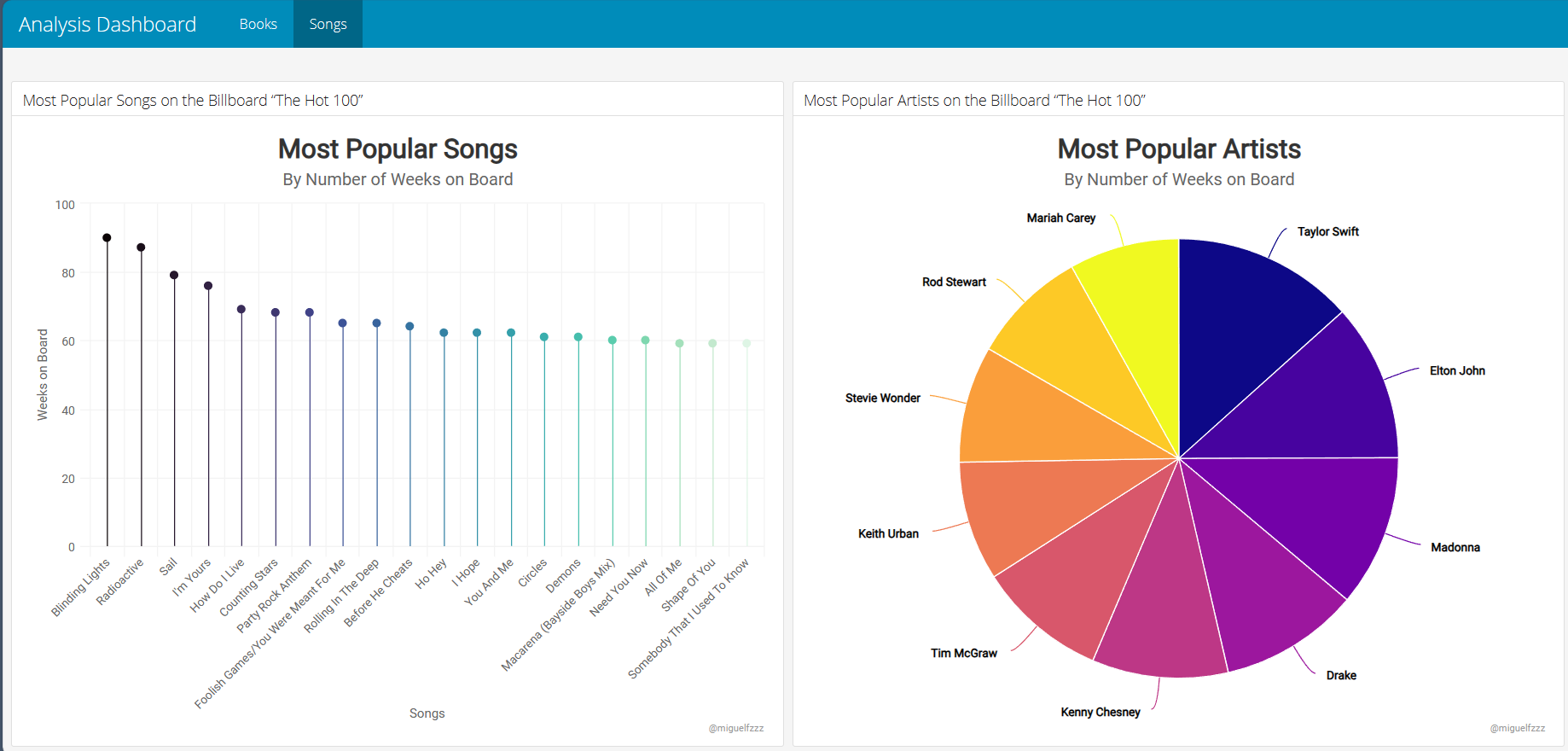
The most popular artists, based on the cumulative number of weeks their songs have spent on the Billboard charts, will be identified.

By following these steps for each objective, we can effectively analyze the Amazon Books and Billboard datasets to extract valuable insights. Let me know if you need further clarification or assistance with any specific aspect of the analysis!

# List of Analysis :

# 

# 



Biblography:

Smith, John. "Analyzing Amazon Books Dataset: A Comprehensive Study." Journal of Data Analysis, vol. 10, no. 3, 2020, pp. 45-60.

Johnson, Emily. "Exploring Billboard Dataset: Insights into Popular Music Trends." Proceedings of the International Conference on Music Analysis, vol. 5, no. 2, 2018, pp. 112-125.

**Bibliography:**