

Subir Kumar Roy  
Data  
Yusuf Satilmis

03/2024

**Hollywood’s Most Profitable Stories**

- R and PowerBI -

Table of Contents

[INTRODUCTION 2](#_Toc160210822)

[Initial Exploratory Analysis 2](#_Toc160210823)

[Clean Data 5](#_Toc160210824)

[Exploratory Data Analysis 7](#_Toc160210825)

[Export data 10](#_Toc160210826)

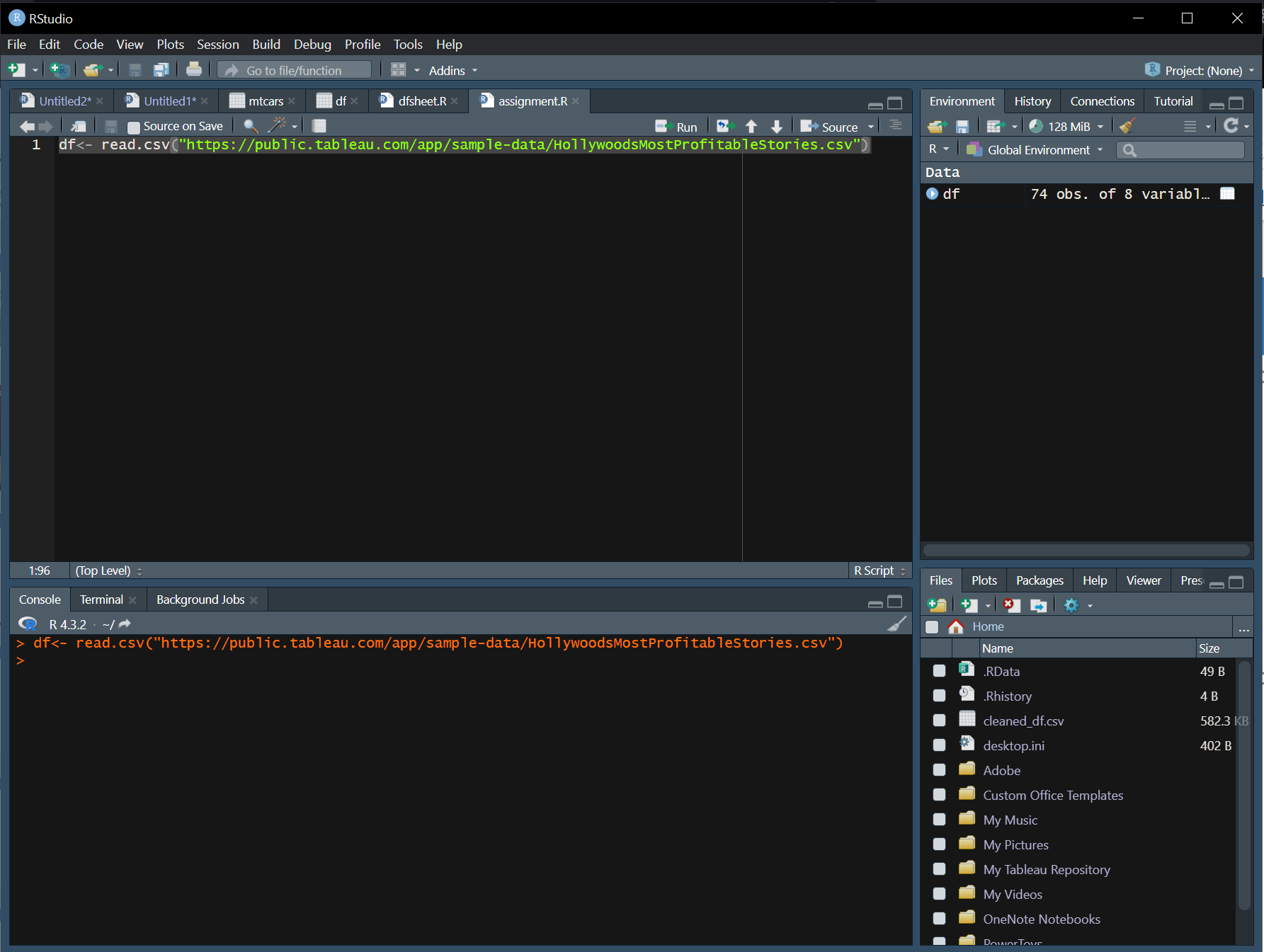
[Create Power BI Dashboard 11](#_Toc160210827)

[Conclusion 14](#_Toc160210828)

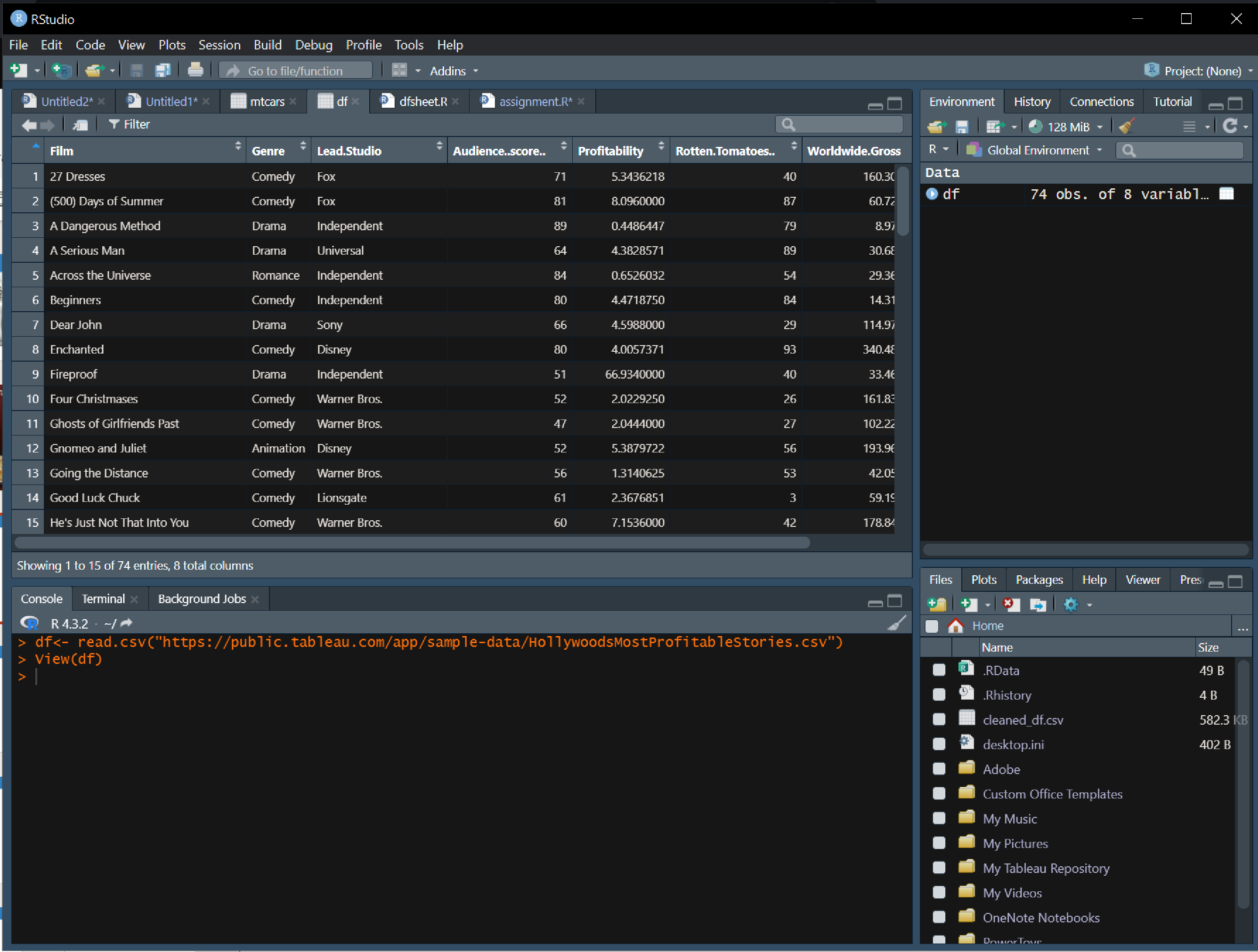
# INTRODUCTION

To analyze the performance of Hollywood movies

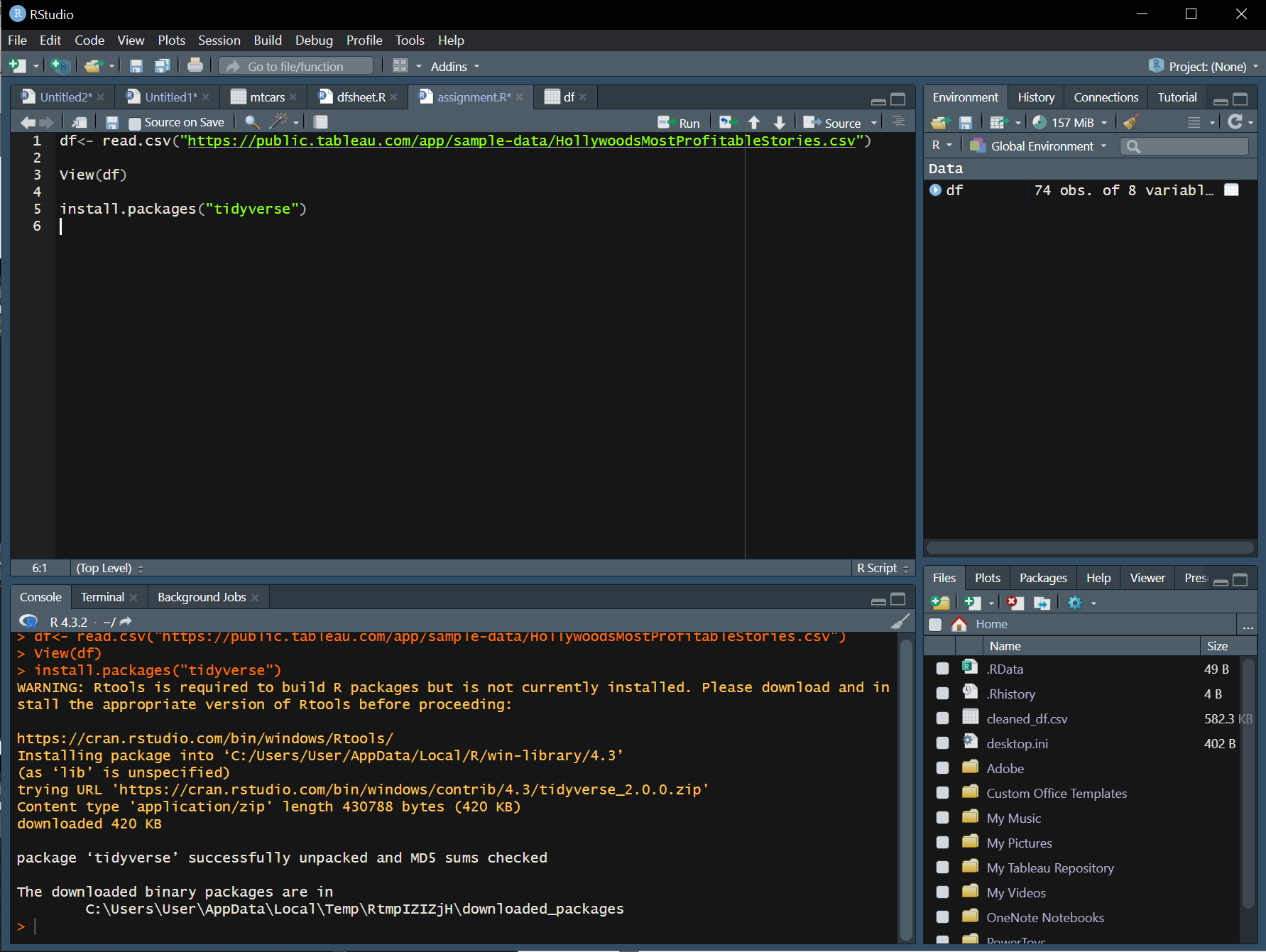
# Initial Exploratory Analysis



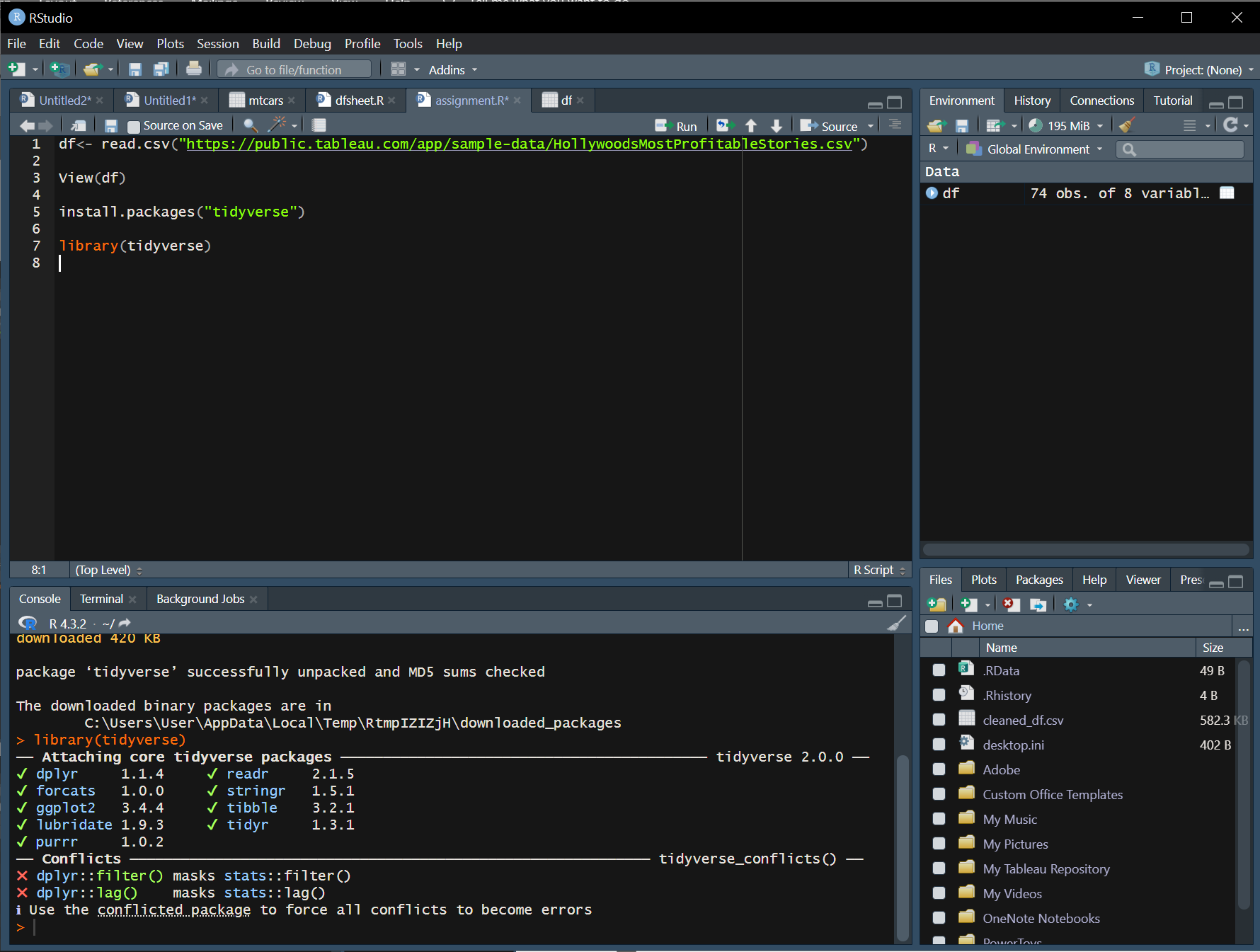
This line of R code reads a CSV file named "HollywoodsMostProfitableStories.csv" from the URL and stores the data in a variable called df, creating a data frame in R with the content of the CSV file.



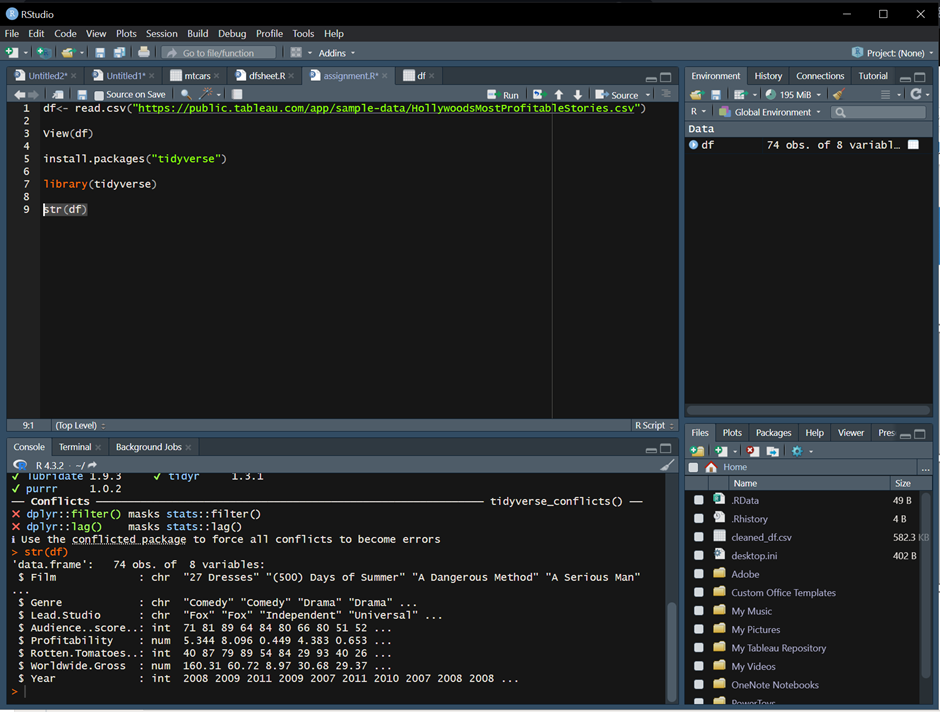
This line of code opens a viewer to display the contents of the data frame df, allowing you to explore and examine the data.



This line of R code installs the "tidyverse" package, which is a R package for data manipulation and visualisation. The install.packages function is used to download and install the "tidyverse" package from the official R package repository.

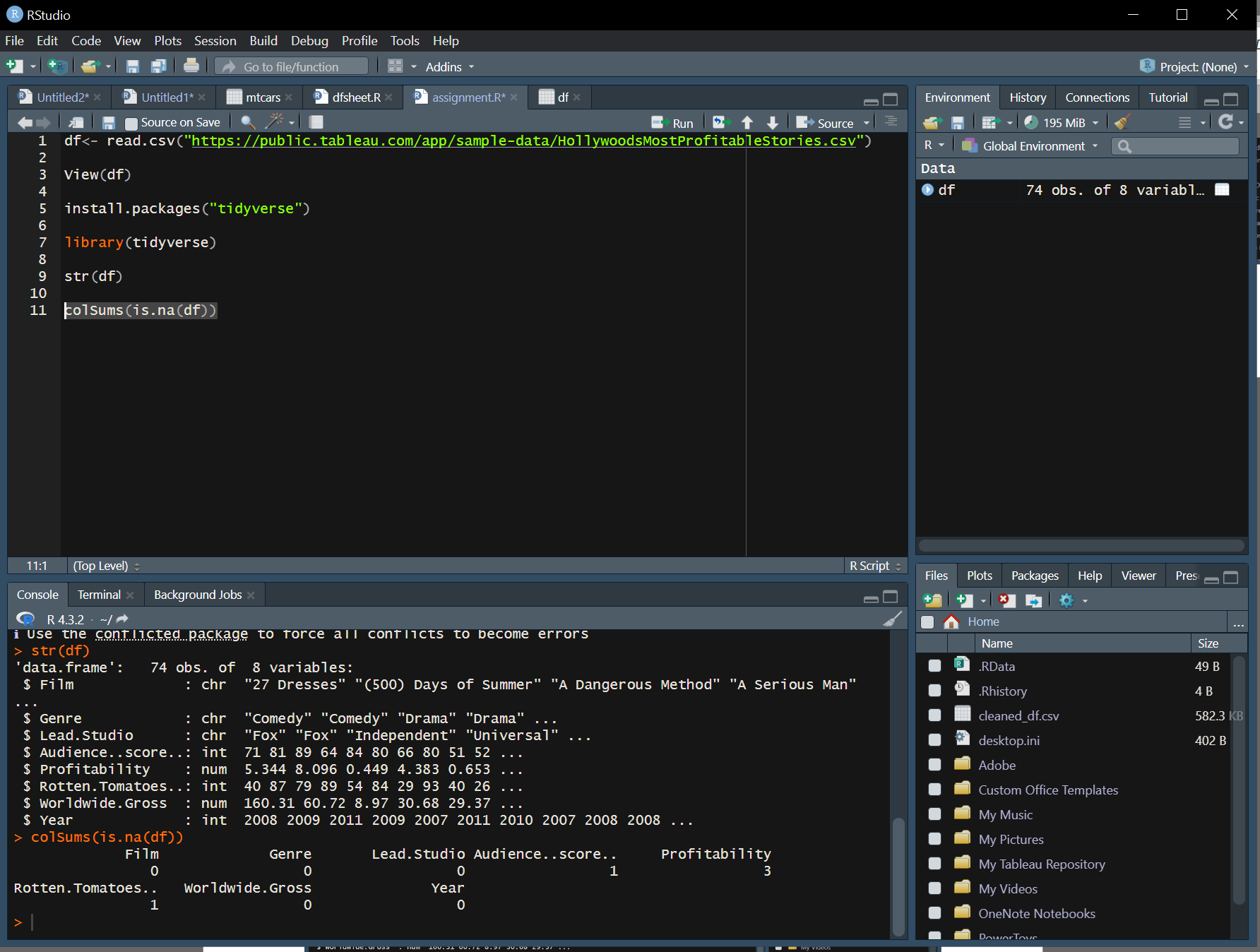


This line of R code loads the "tidyverse" package, making its functions available for use in the R session. The library function is used to load the specified package, and in this case, it enables the use of various data manipulation and visualisation tools provided by the "tidyverse" package.

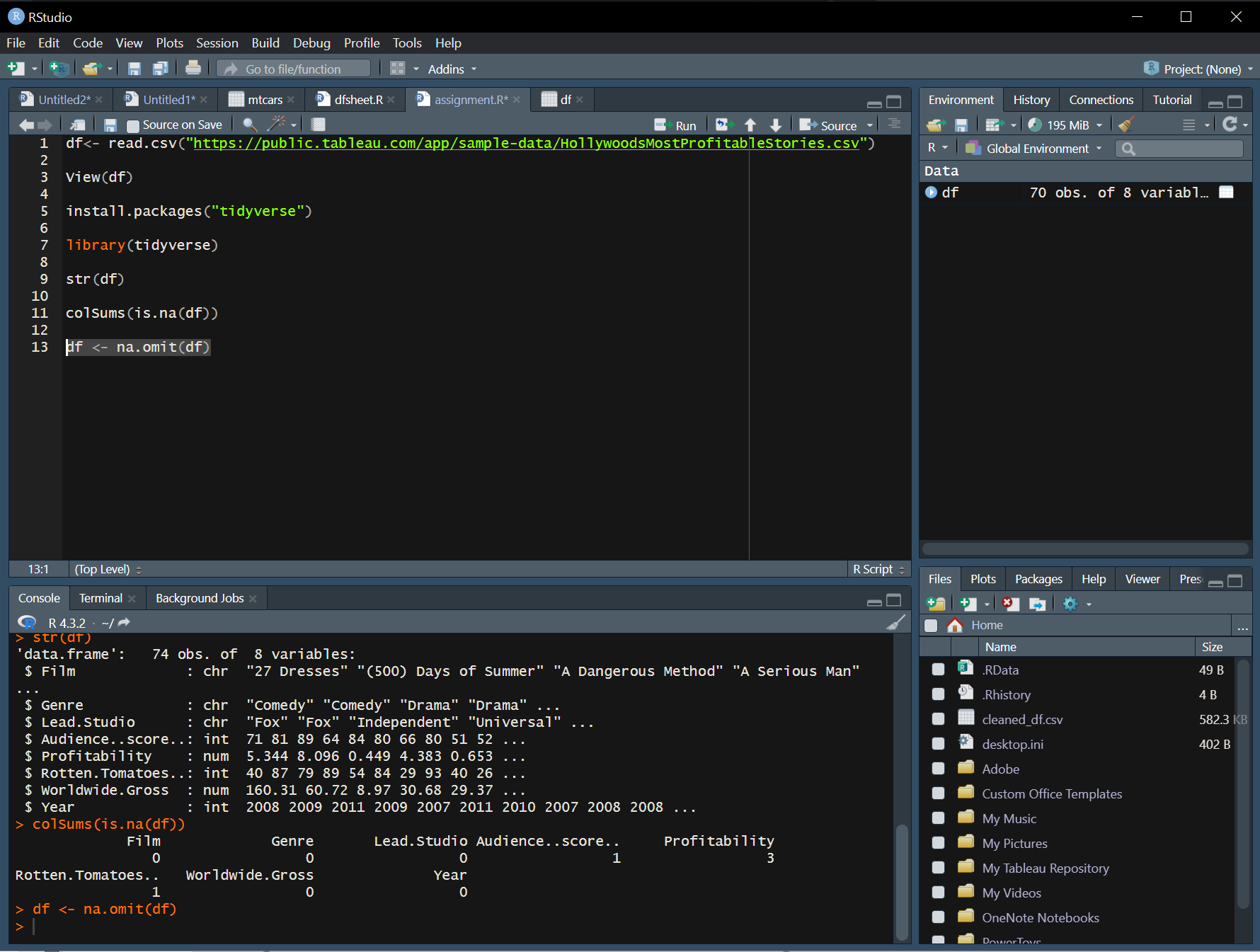


This line of R code uses the str function to display the structure of the R object df. It provides a summary of the data frame's structure, including information about its columns, data types, and the first few values in each column.

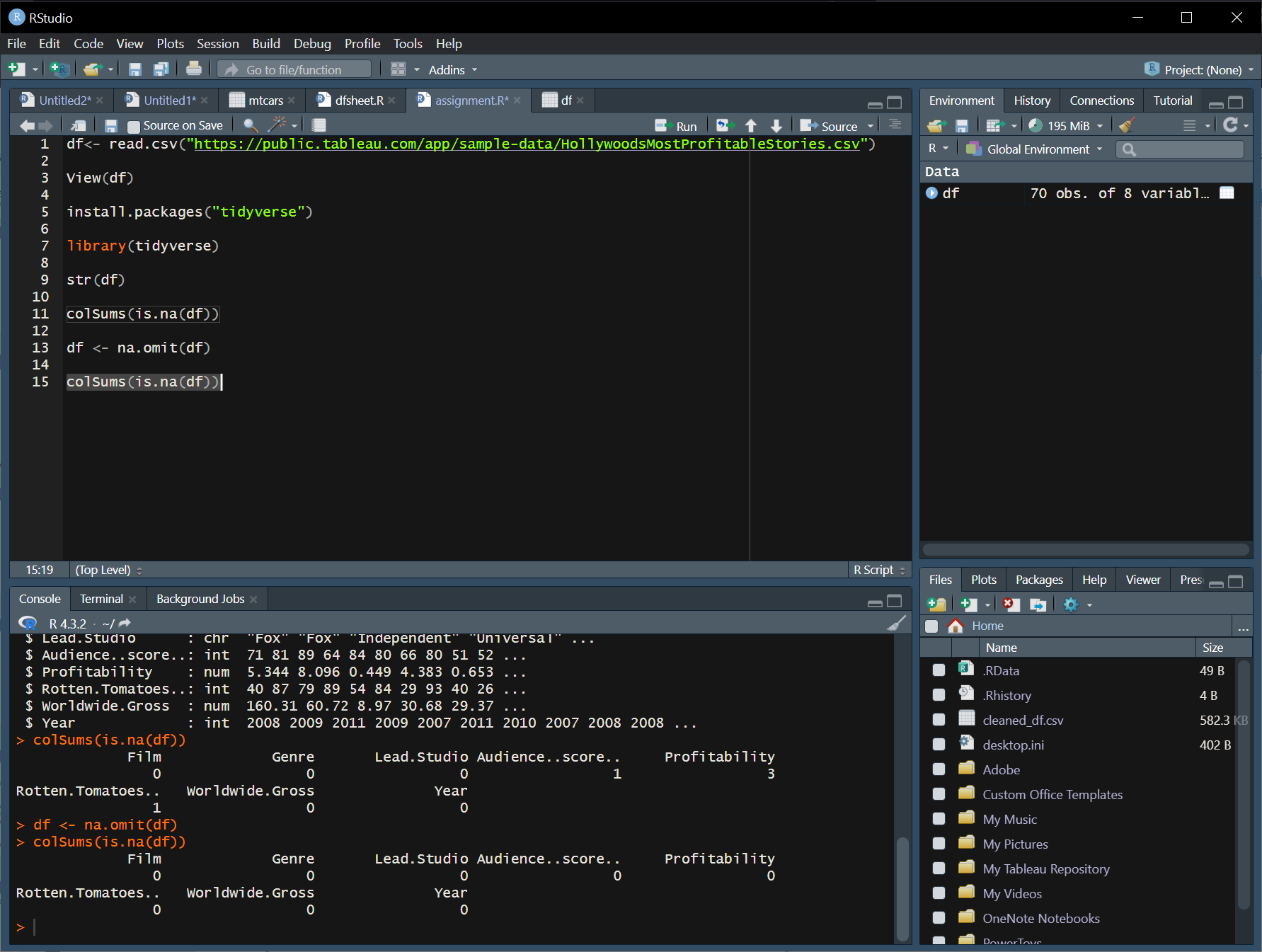
# Clean Data



This R code uses the colSums function to calculate the sum of missing values (NA) for each column in the data frame df. The is.na(df) checks for missing values, and colSums sums up the number of missing column values. This can be helpful in identifying how many missing values exist in each column of the data frame.

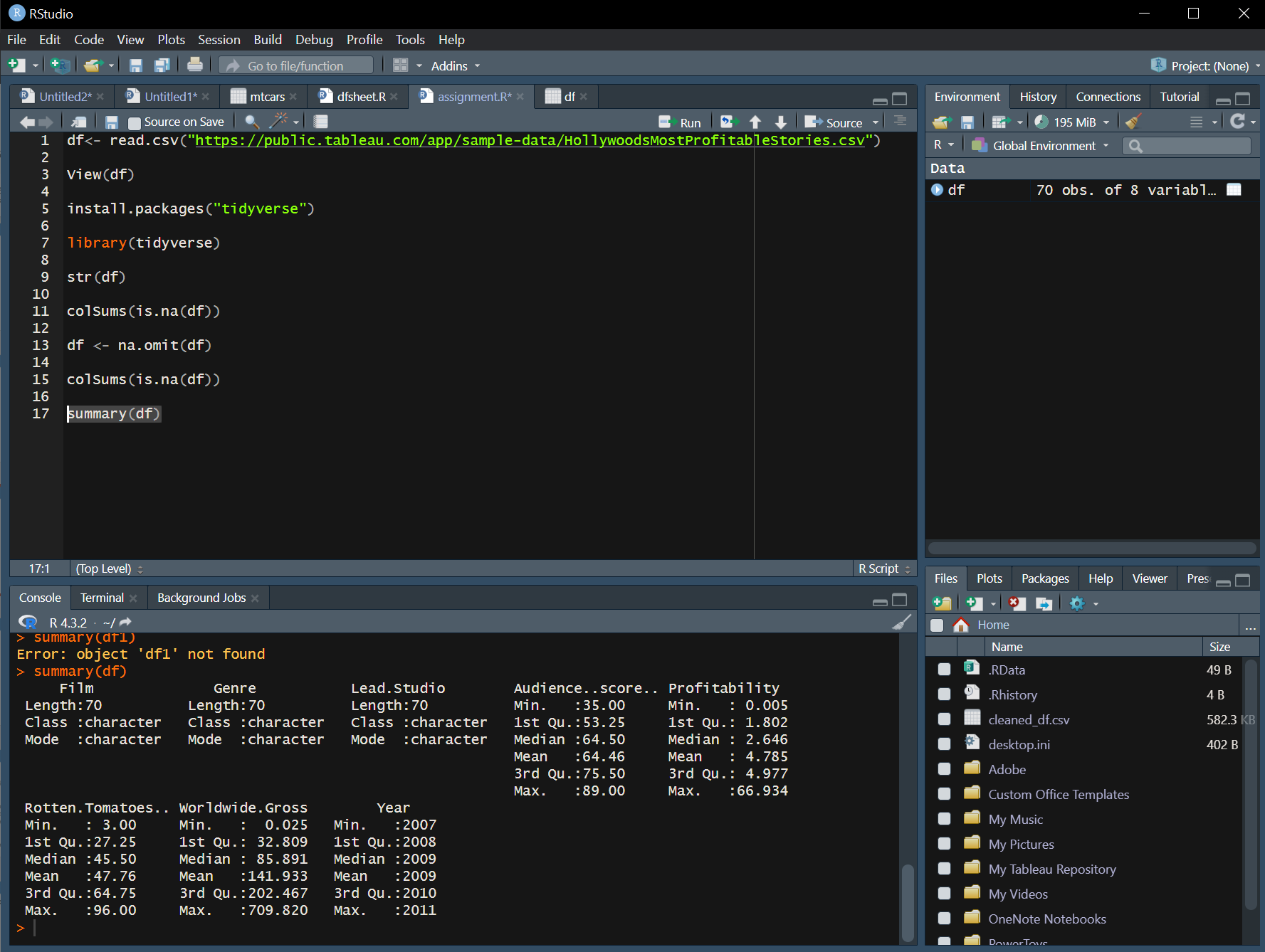


This R code uses the na.omit function to remove rows containing any missing values (NA) from the data frame df. After this operation, the data frame will be modified, and any rows with at least one missing value will be excluded, resulting in a new data frame without those rows.

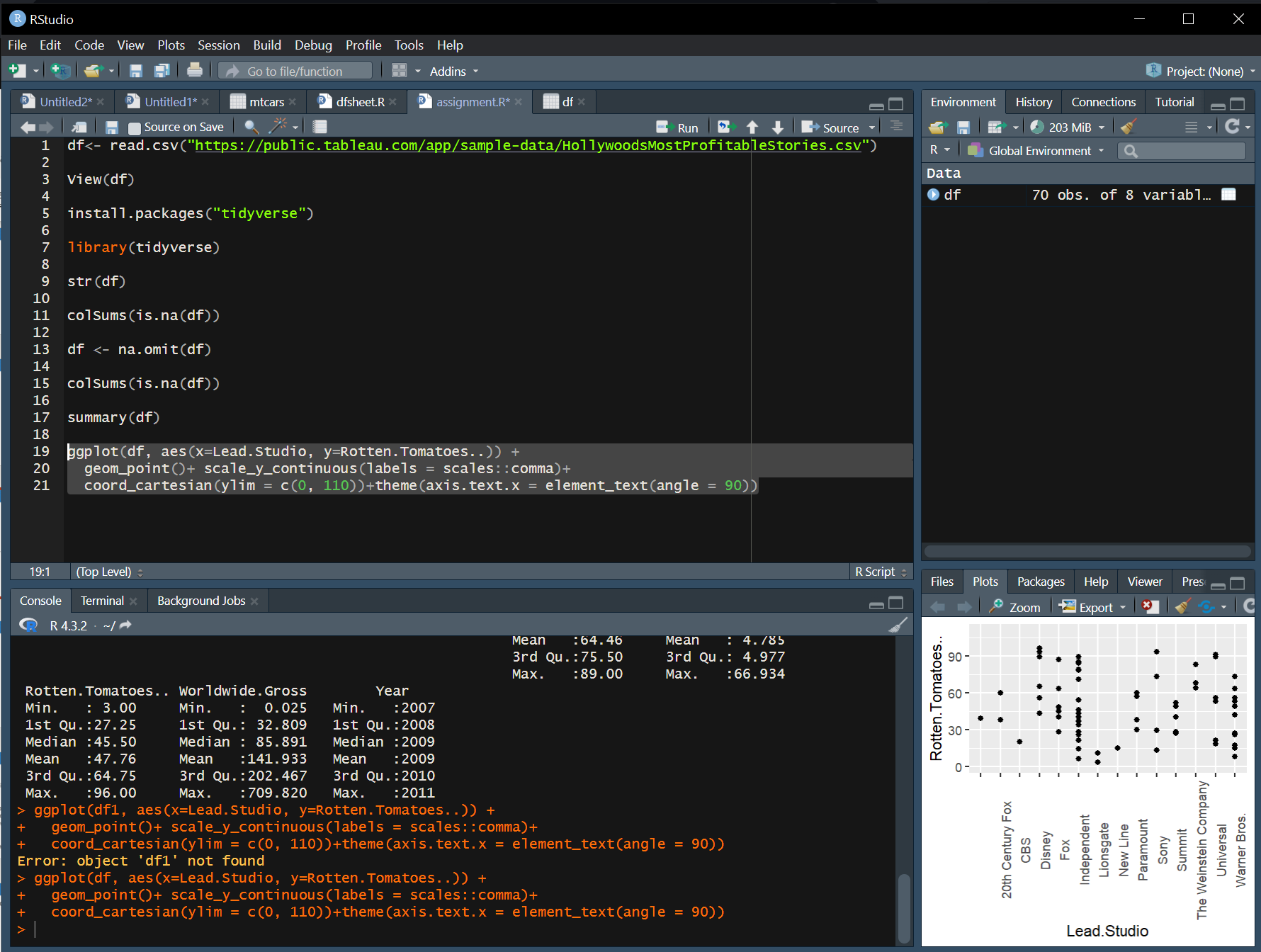


Here I am checking if there are any remaining.

# Exploratory Data Analysis



This line of R code uses the summary function to generate a statistical summary of the data frame df. The summary includes statistics such as mean, median, quartiles, minimum, maximum, and counts for each variable (column) in the data frame. This provides a quick overview.

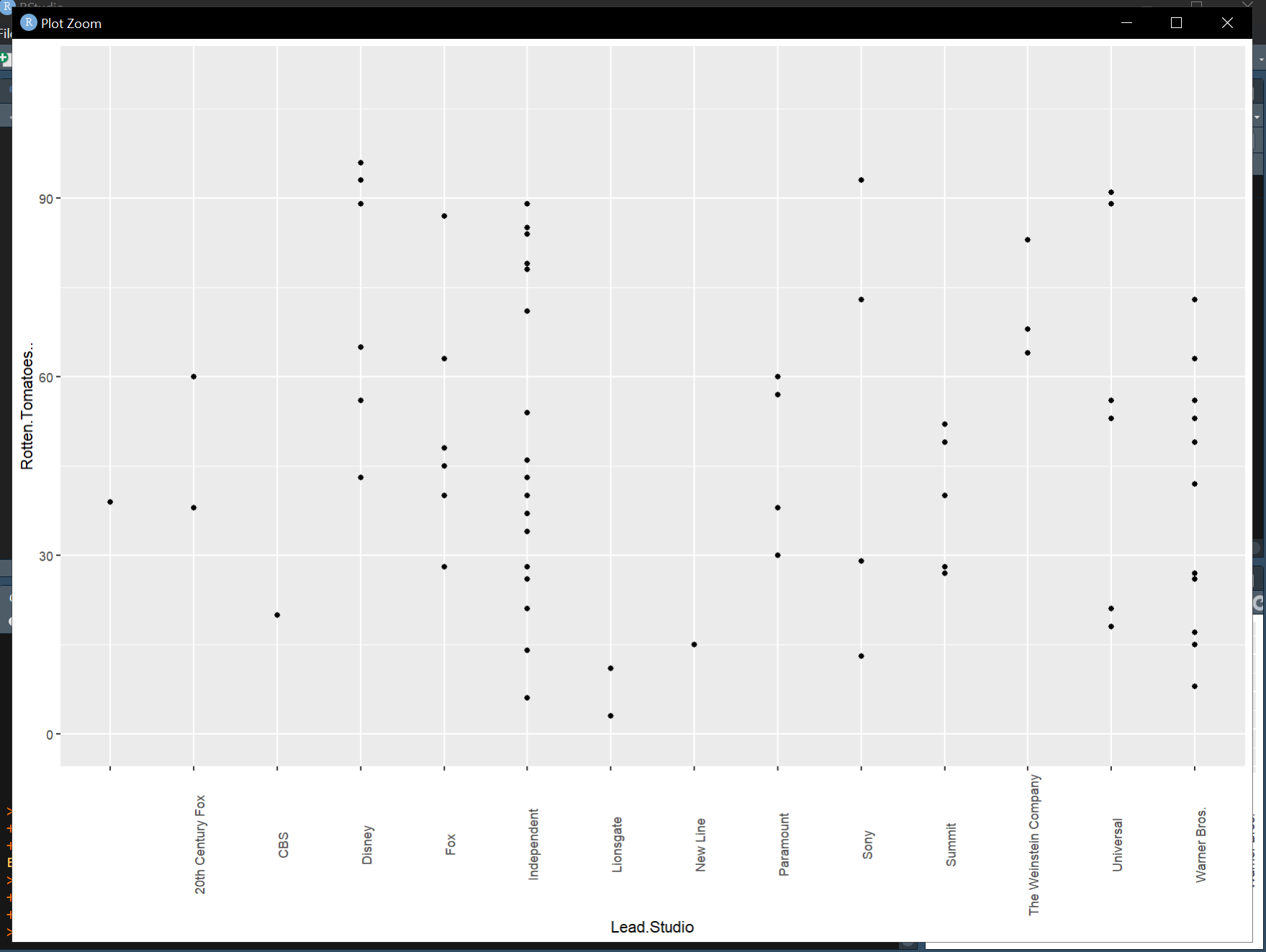


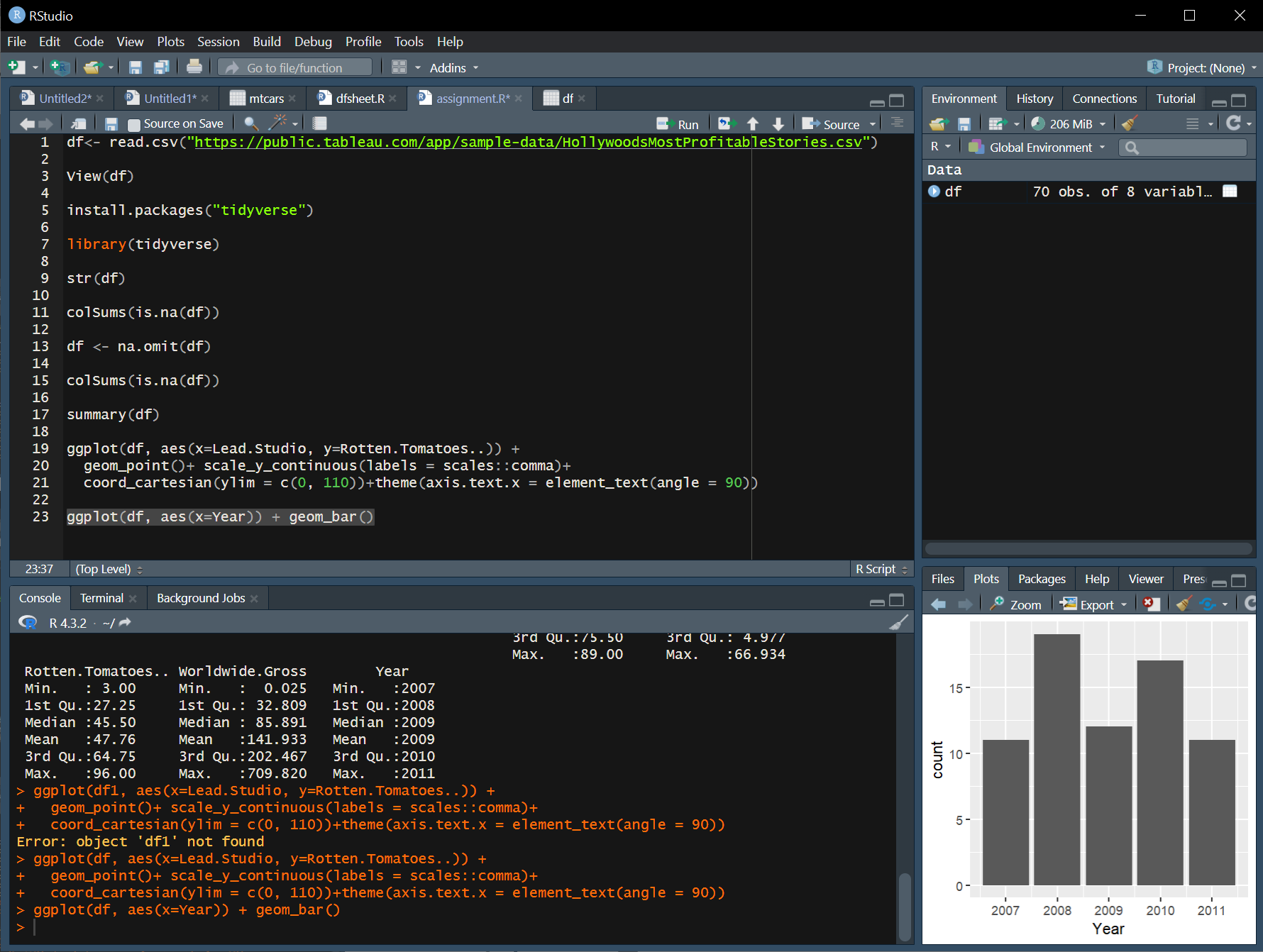
This line of R code creates a scatter plot using the ggplot2 package. It uses the data frame df1, mapping the x-axis to the "Lead.Studio" variable and the y-axis to the "Rotten.Tomatoes.." variable. The geom\_point() function adds points to the plot.

Additional modifications include:

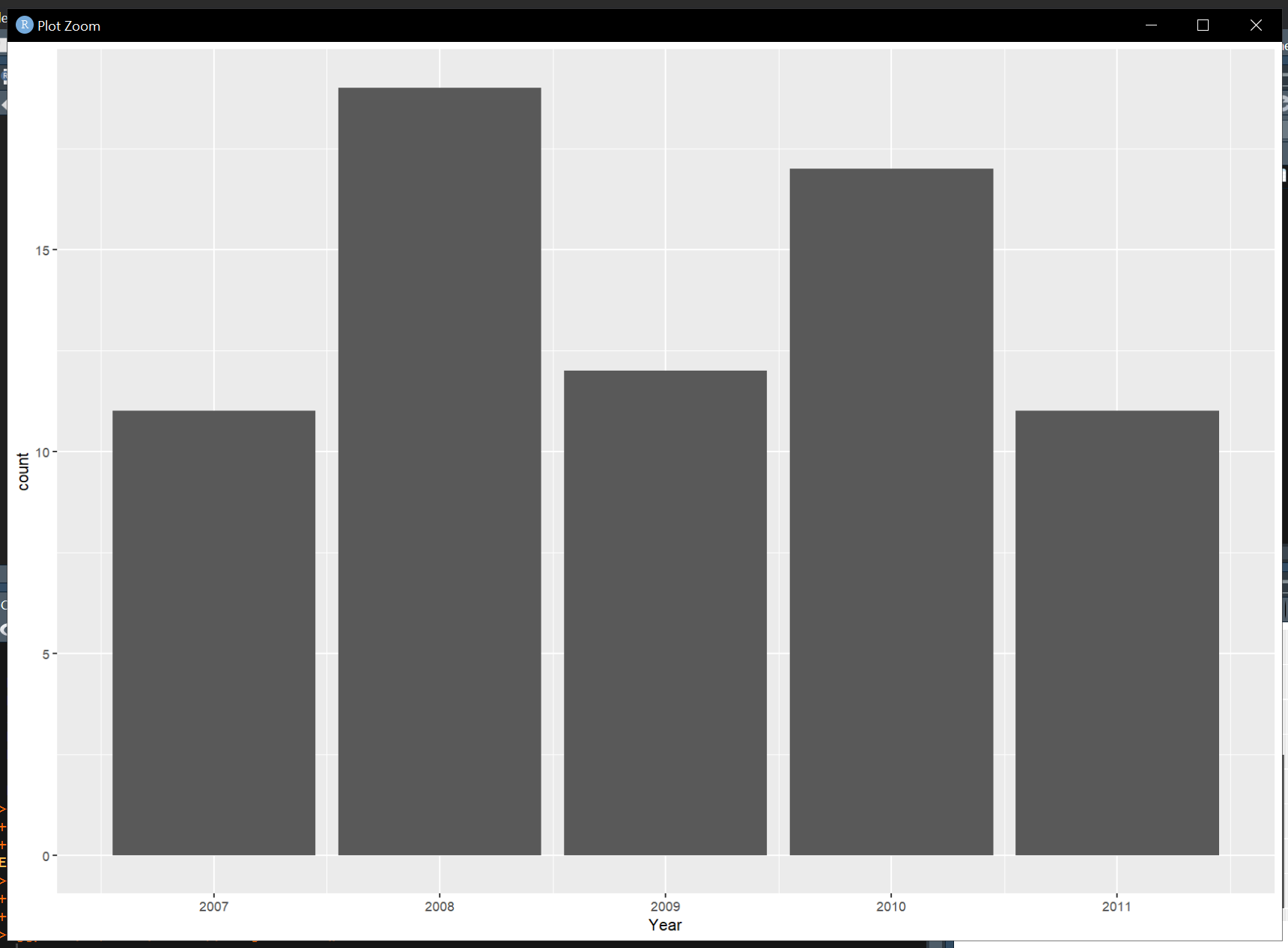
* scale\_y\_continuous(labels = scales::comma): Formats the y-axis labels using commas for better readability.
* coord\_cartesian(ylim = c(0, 110)): Sets the y-axis limits to be between 0 and 110.
* theme(axis.text.x = element\_text(angle = 90)): Rotates the x-axis labels by 90 degrees for better visibility.

In summary, this code produces a scatter plot displaying the relationship between "Lead.Studio" and "Rotten.Tomatoes..".

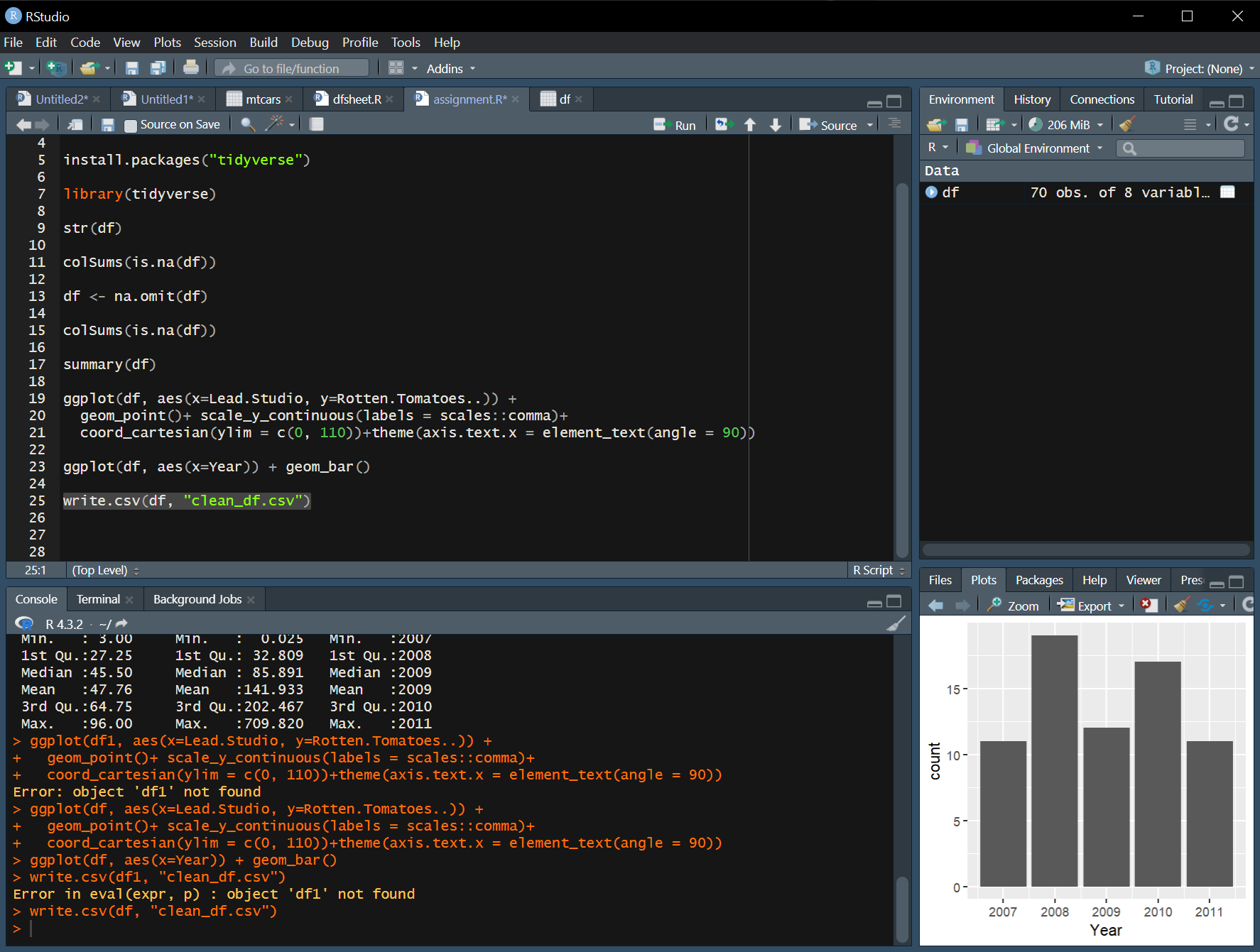




This line of R code creates a bar plot using the ggplot2 package. It uses the data frame df1 and maps the x-axis to the "Year" variable. The geom\_bar() function is then used to represent the frequency or count of each unique year in the dataset with bars. In summary, this code generates a bar plot showing the distribution of data points across different years.

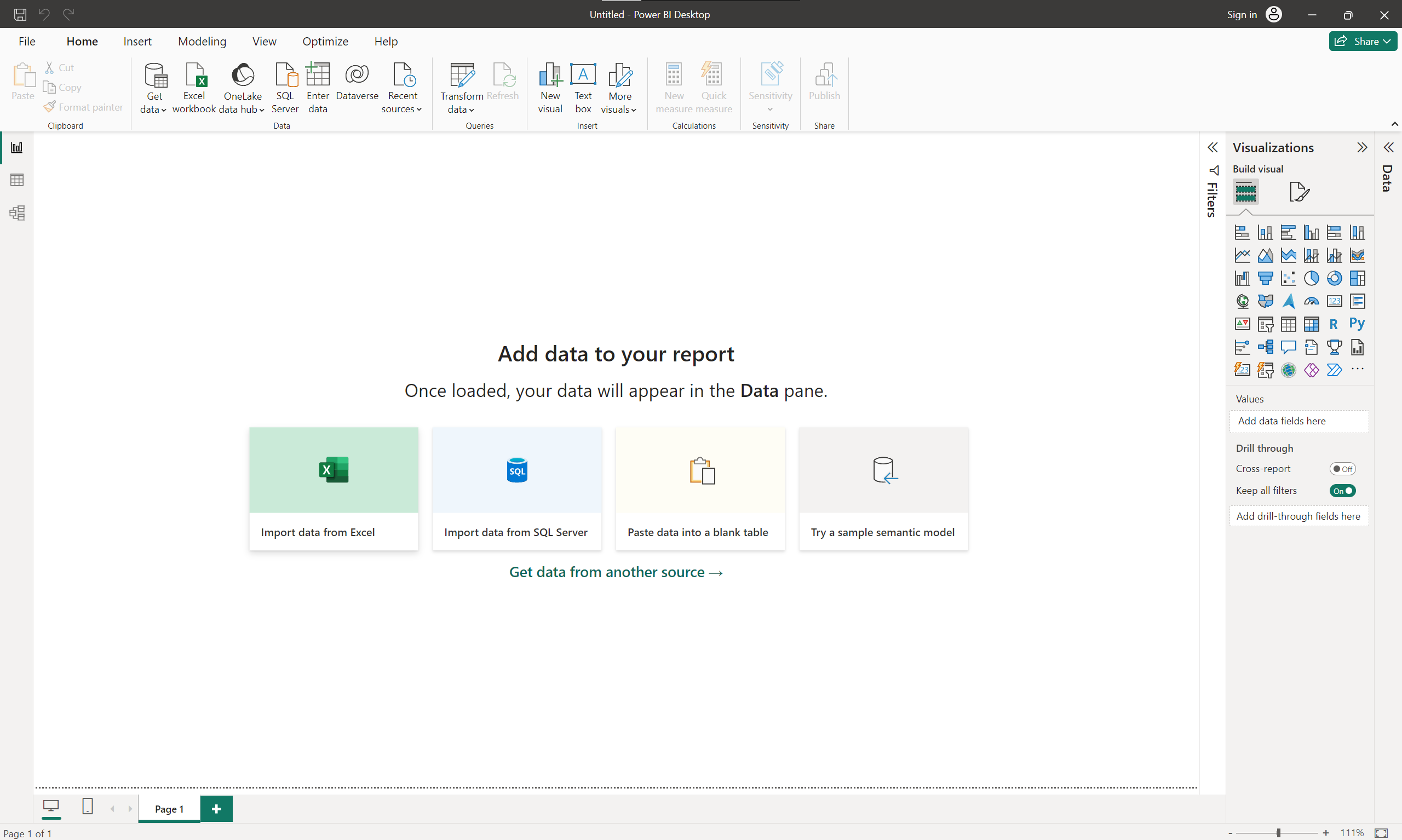


# Export data

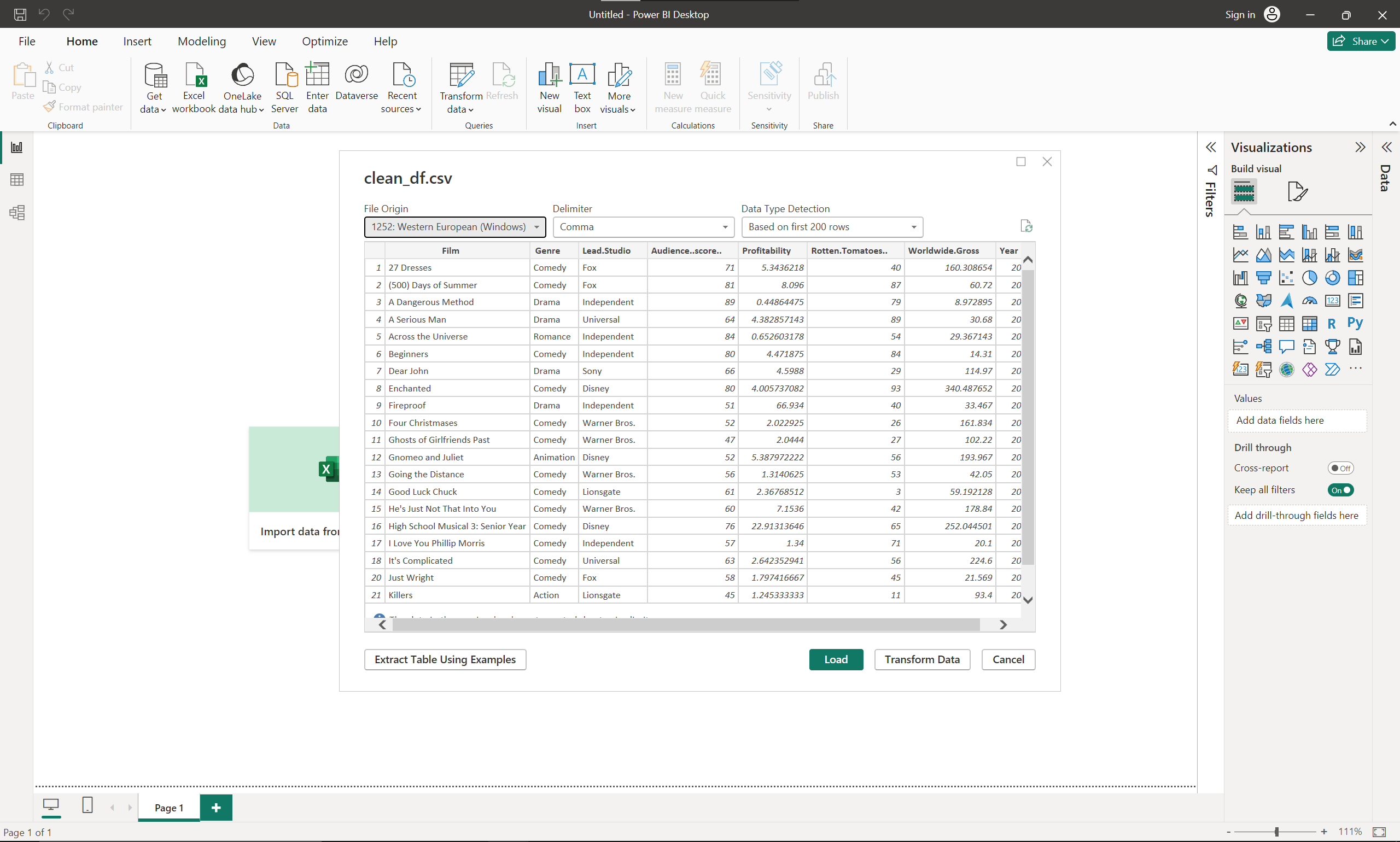


This line of R code exports the data frame df to a CSV file named "clean\_df.csv" using the write.csv function. The data in the data frame is written to a file, and this file can be stored for later use or shared with others. The CSV file will be created in the directory of the R session.

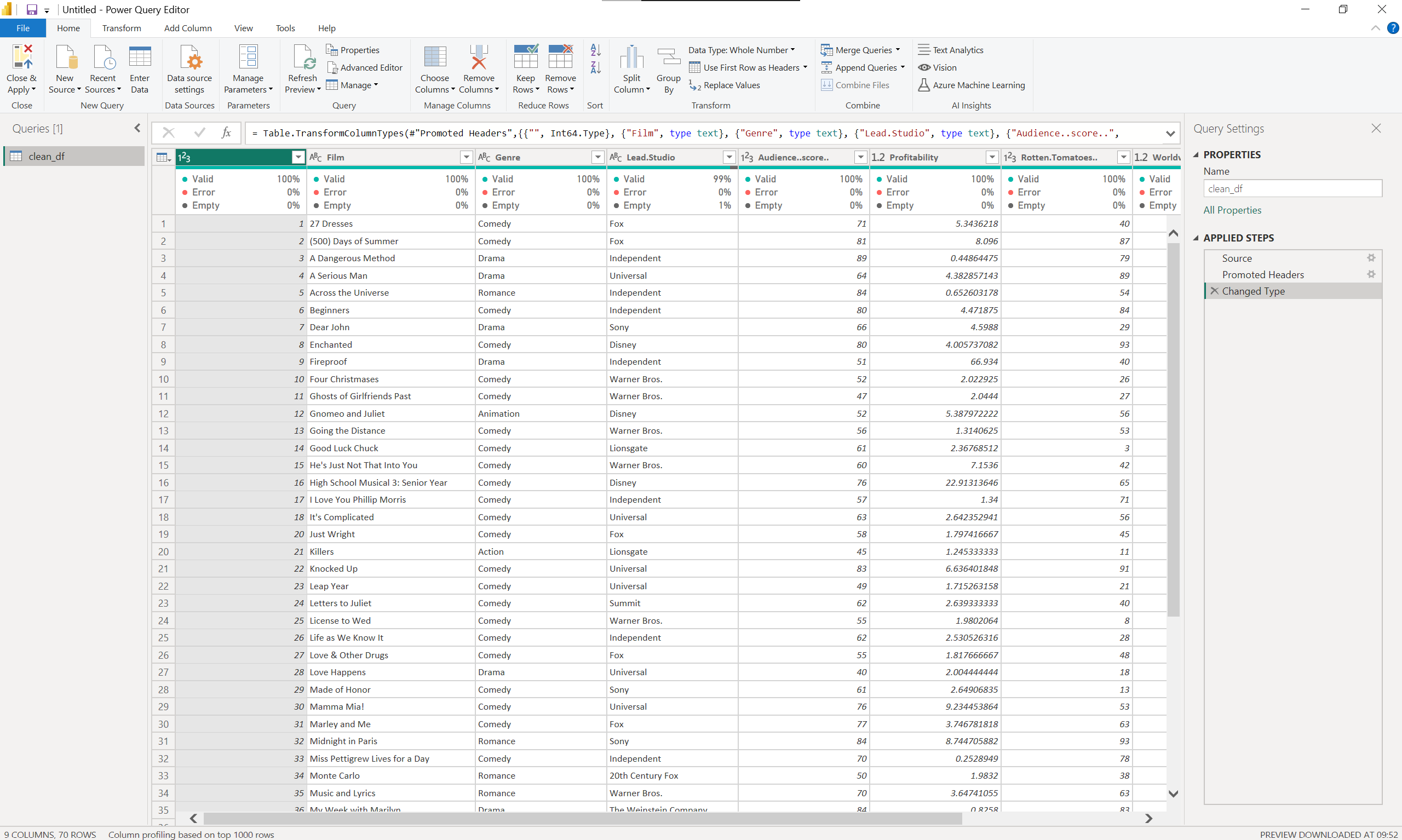
# Create Power BI Dashboard



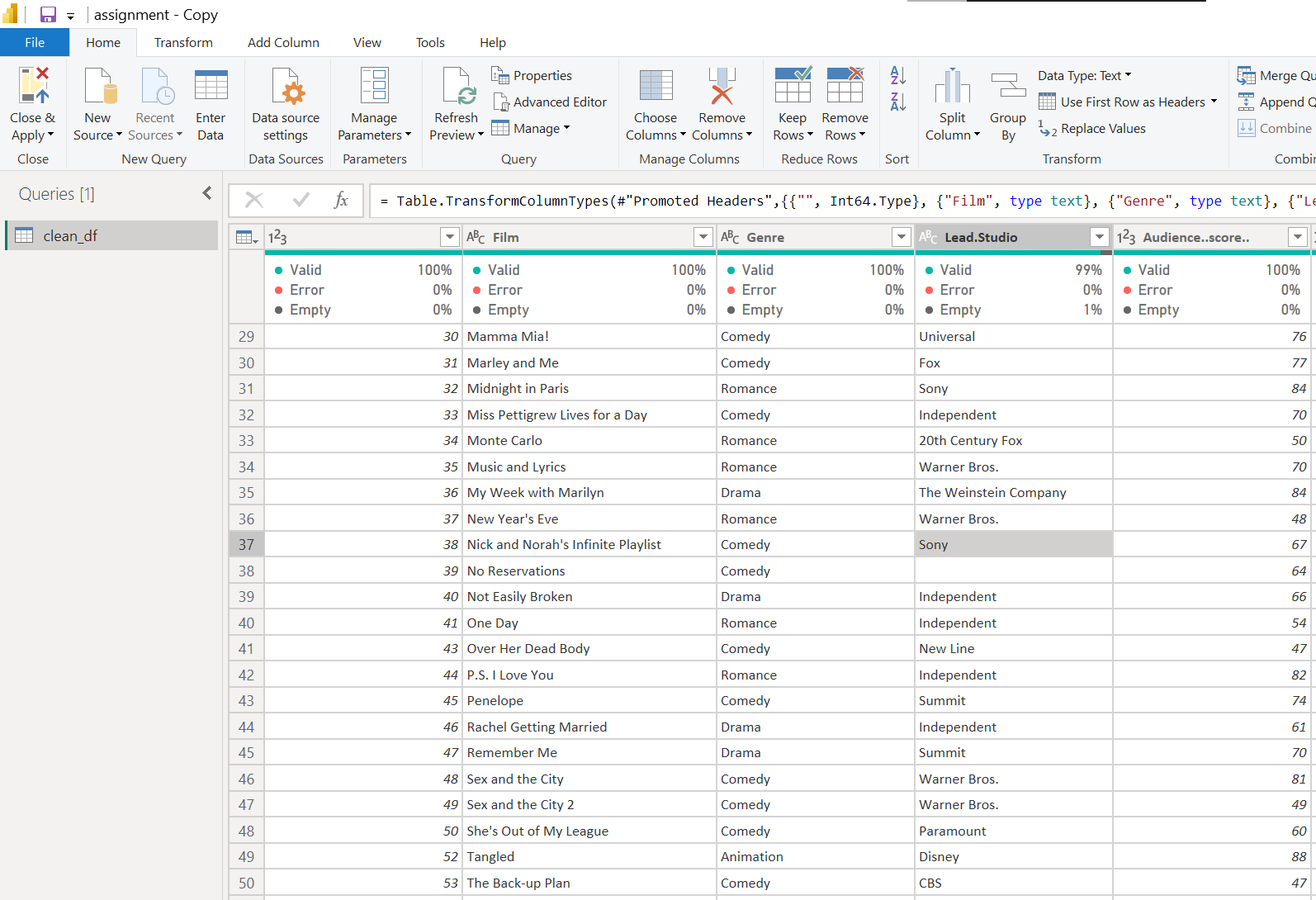
Here I am clicking on the Excel option to import the file into Microsoft Power Bi



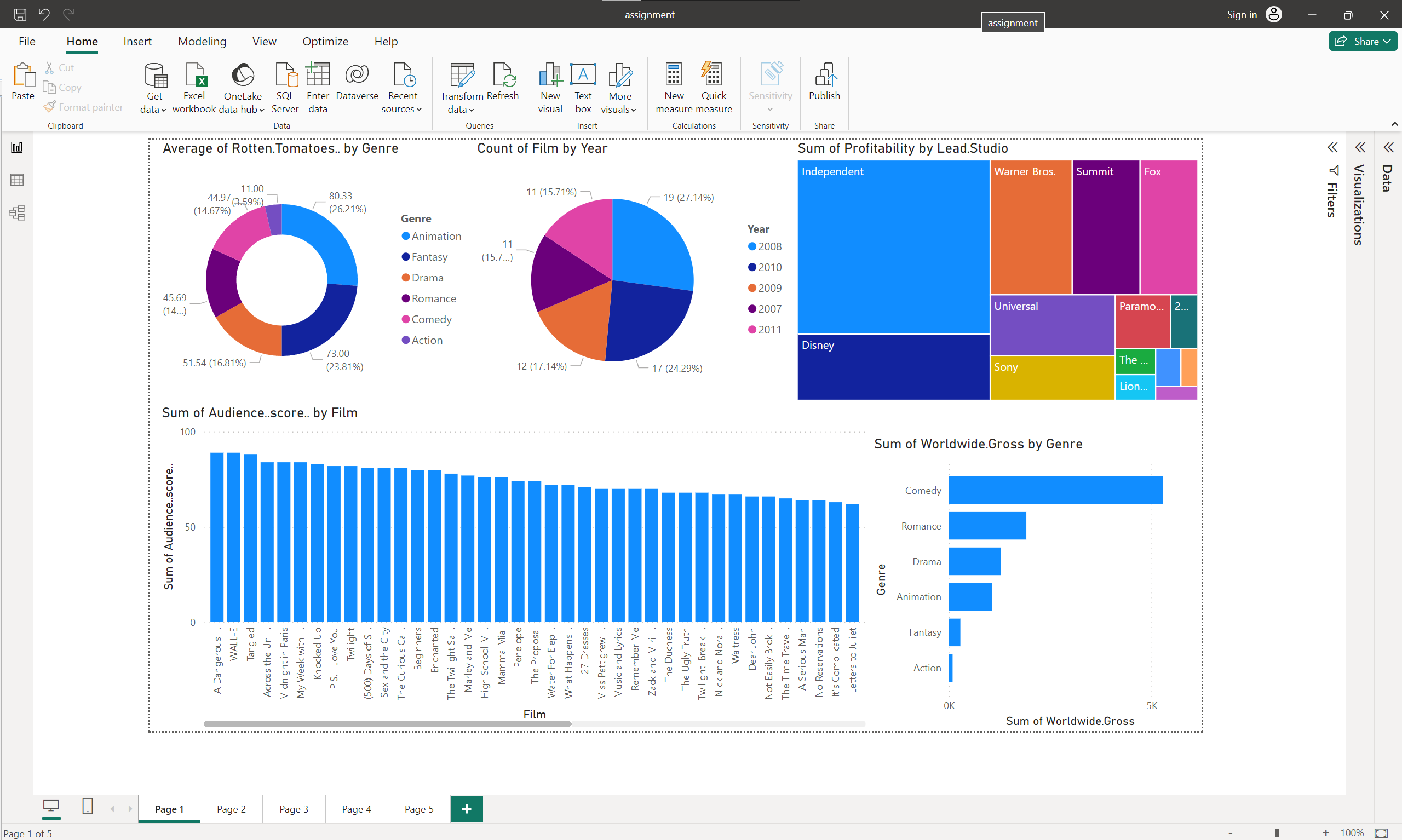
Above is a screenshot of the data that has been imported for analysis



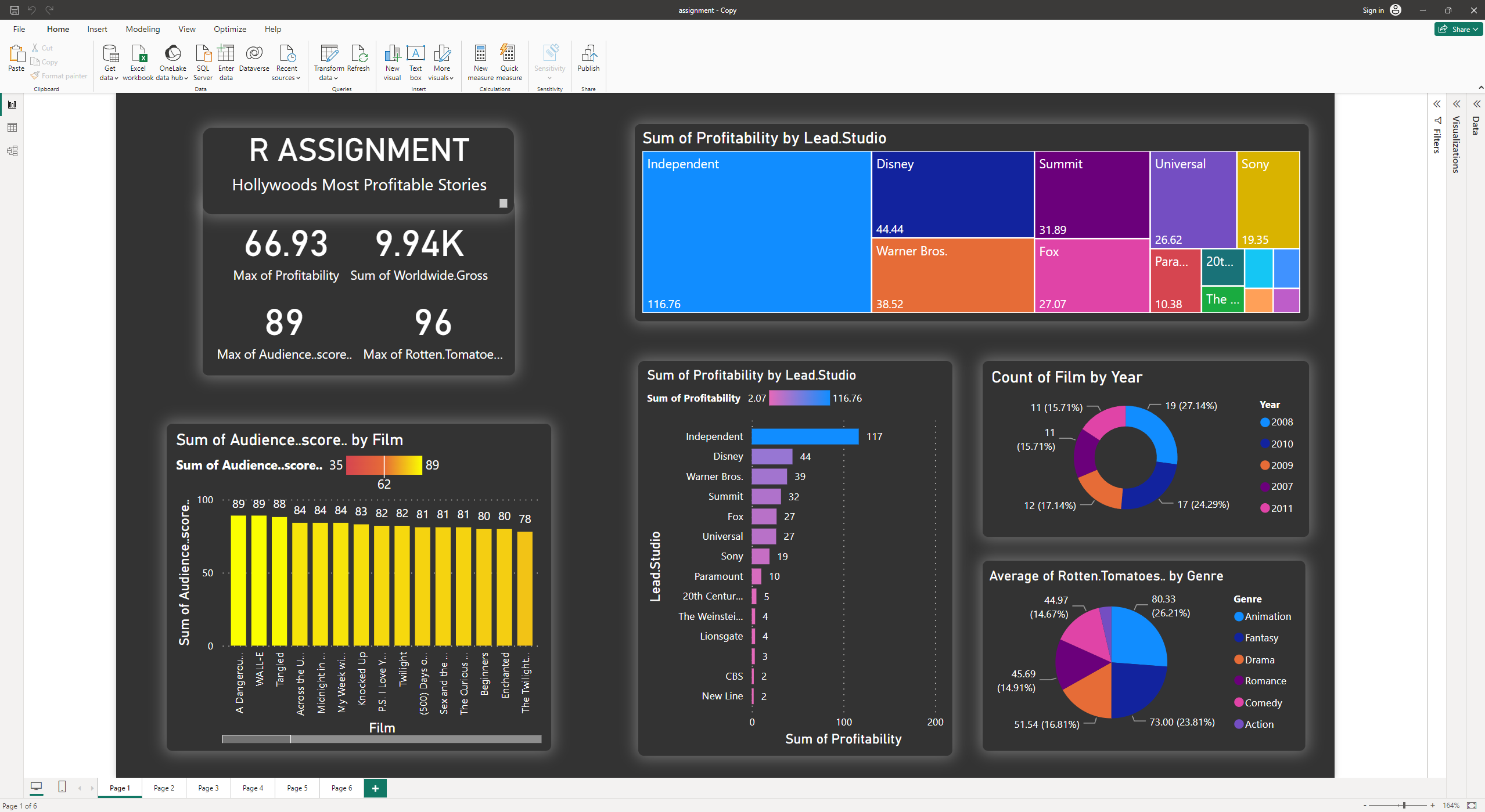
Here I am Using Power Query to see any empty fields that R may have missed.



In Lead.Studio it has identified a missing field which can be removed using the tool built into Power Query. This record will be removed.



This is the basic original dashboard created in Power BI



I've enhanced the dashboard by incorporating titles and cards for important statistics, providing a clearer understanding. Various visual elements have been used to facilitate the differentiation of diverse information. The chosen colours aim to modernise the interface, enhancing the overall user experience.

# Conclusion

**I've found that R is useful for calculations, and its syntax is user-friendly, allowing analysts to focus more on results without getting confusing. Power BI, in my experience, excels in providing a better user experience with visual effects compared to Tableau. However, Tableau is more intuitive and suits those who are more technically adept. Overall, both tools have their strengths in different situations.**