Module 5 R Practice

Ayush Anand

NORTHEASTERN UNIVERSITY:
COLLEGE OF PROFESSIONAL STUDIES
ALY 6010: PROBABILITY THEORY AND INTRODUCTORY STATISTICS
PROFESSOR XYZ
OCTOBER 25TH, 2024

About the Dataset:

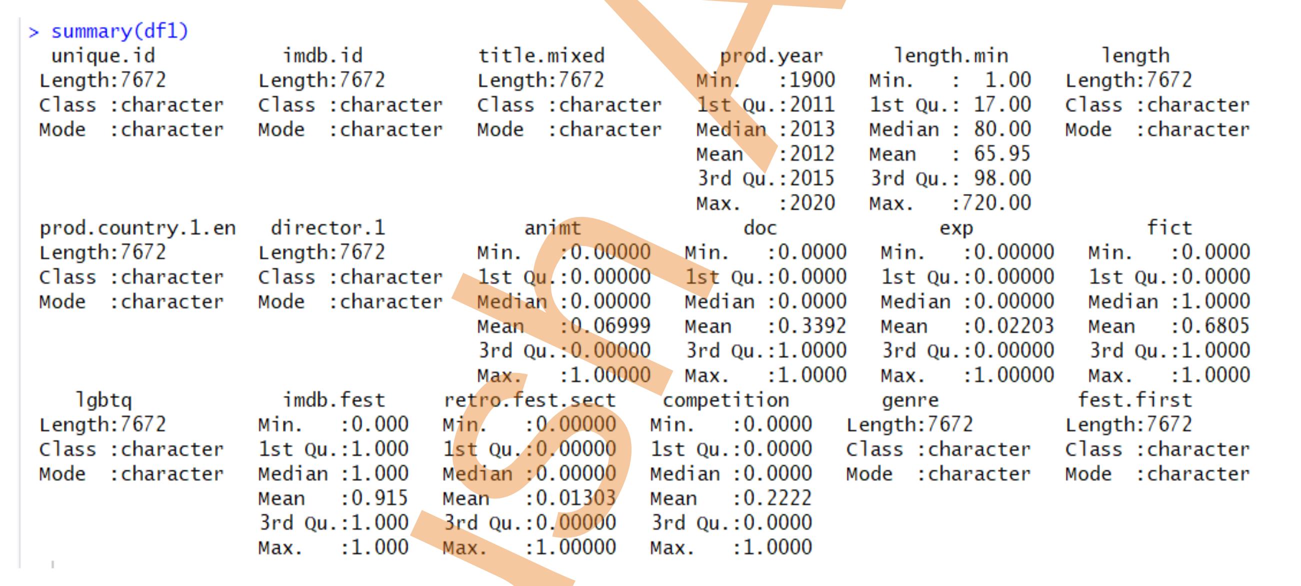
The dataset (1_film-dataset_festival-program_wide.csv) presents a comprehensive overview of 9,348 unique films. Each film is listed once, with information about the first festival it appeared at. This dataset includes essential details like IMDb IDs, titles, release years, genres, directors, and festival-related data. Although it's more concise, this dataset provides valuable insights into festival participation and film characteristics, making it suitable for various analyses.
These datasets are a valuable resource for studying the film industry's involvement in international festivals, production trends, and the global film festival circuit.

Introduction

The aim of this analysis is to explore relationships between variables in a dataset related to film festivals. The analytical question guiding this study is: Does the year of film production significantly influence its participation in a retrospective section? By examining this question, we seek to determine whether variables such as production year, festival participation, and film characteristics can be used to predict the retrospective inclusion of a film. To address this, a sequence of statistical methods was employed: preliminary data cleaning, descriptive analysis, hypothesis testing, correlation analysis, and linear regression modeling. Each step aimed to clarify the relationships between variables and uncover any predictive relationships.

Initial Exploration:

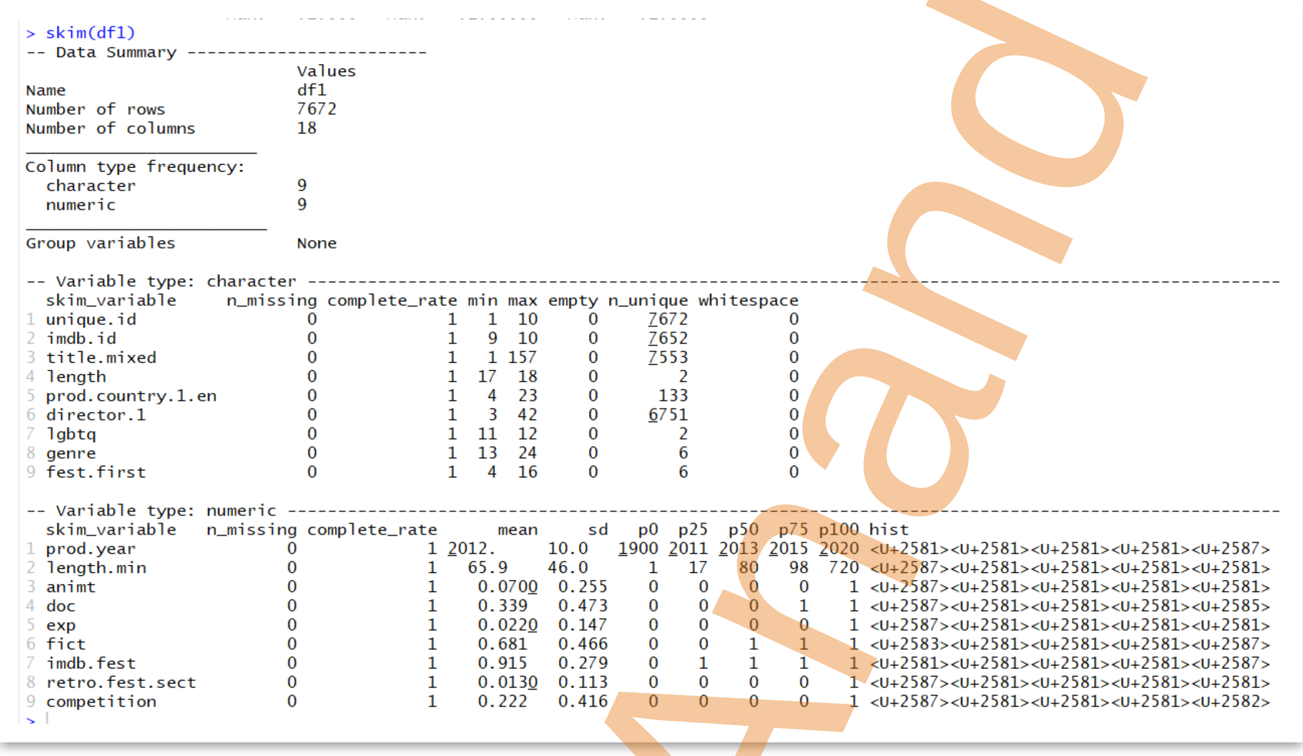
Used head() to preview the first few rows of the dataset. Utilized str() to inspect the structure of the dataset and understand the data types and variables. Summarized the dataset with summary() to gather descriptive statistics.



```
> summary(df1)
  unique.id           imdb.id          title.mixed          prod.year      length.min        length         
 Length:7672        Length:7672        Length:7672        Min.   :1900   Min.   :  1.00   Length:7672       
 Class :character   Class :character   Class :character   1st Qu.:2011   1st Qu.: 17.00   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Median :2013   Median : 80.00   Mode  :character  
                                                          Mean   :2012   Mean   : 65.95                     
                                                          3rd Qu.:2015   3rd Qu.: 98.00                     
                                                          Max.   :2020   Max.   :720.00                     
 prod.country.1.en   director.1            animt              doc              exp               fict       
 Length:7672        Length:7672        Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000  
 Class :character   Class :character   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000  
 Mode  :character   Mode  :character   Median :0.00000   Median :0.0000   Median :0.00000   Median :1.0000  
                                       Mean   :0.06999   Mean   :0.3392   Mean   :0.02203   Mean   :0.6805  
                                       3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000  
                                       Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000  
    lgbtq             imdb.fest     retro.fest.sect     competition        genre            fest.first       
 Length:7672        Min.   :0.000   Min.   :0.00000   Min.   :0.0000   Length:7672        Length:7672       
 Class :character   1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.0000   Class :character   Class :character  
 Mode  :character   Median :1.000   Median :0.00000   Median :0.0000   Mode  :character   Mode  :character  
                    Mean   :0.915   Mean   :0.01303   Mean   :0.2222                                        
                    3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.0000                                        
                    Max.   :1.000   Max.   :1.00000   Max.   :1.0000                                        
```

Skimming the Dataset:

Applied the **skim()** function for a comprehensive overview of the dataset, providing detailed insights into each variable.

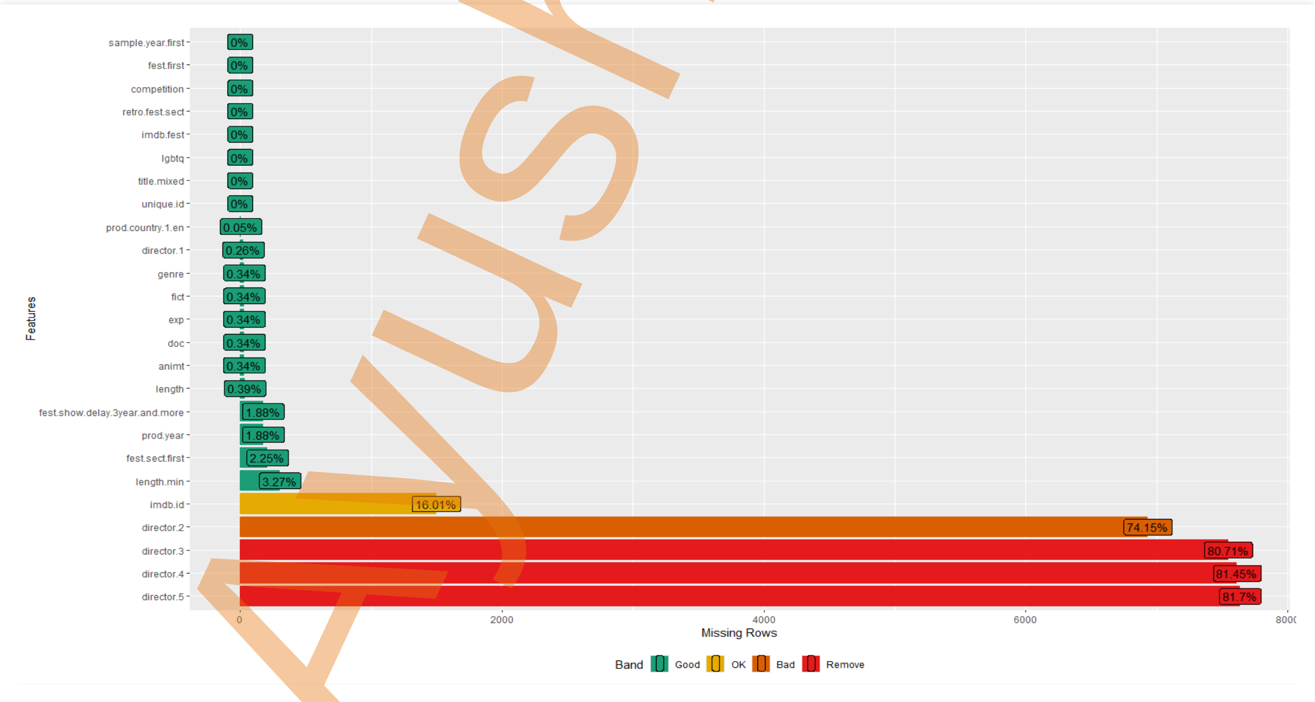


```
> skim(df1)
-- Data Summary ------------------------
                           Values
Name                       df1
Number of rows             7672
Number of columns          18
_______________________
Column type frequency:
  character                9
  numeric                  9
________________________
Group variables            None

-- Variable type: character ------------------------------------------------------------------------------------
  skim_variable     n_missing complete_rate min max empty n_unique whitespace
1 unique.id                 0             1   1  10     0     7672          0
2 imdb.id                   0             1   9  10     0     7652          0
3 title.mixed               0             1   1 157     0     7553          0
4 length                    0             1  17  18     0        2          0
5 prod.country.1.en         0             1   4  23     0      133          0
6 director.1                0             1   3  42     0     6751          0
7 lgbtq                     0             1  11  12     0        2          0
8 genre                     0             1  13  24     0        6          0
9 fest.first                0             1   4  16     0        6          0

-- Variable type: numeric --------------------------------------------------------------------------------------
  skim_variable   n_missing complete_rate      mean     sd   p0  p25  p50  p75 p100 hist
1 prod.year               0             1 2012.     10.0  1900 2011 2013 2015 2020 <U+2581><U+2581><U+2581><U+2581><U+2587>
2 length.min              0             1   65.9    46.0     1   17   80   98  720 <U+2587><U+2581><U+2581><U+2581><U+2581>
3 animt                   0             1    0.0700  0.255   0    0    0    0    1 <U+2587><U+2581><U+2581><U+2581><U+2581>
4 doc                     0             1    0.339   0.473   0    0    0    1    1 <U+2587><U+2581><U+2581><U+2581><U+2585>
5 exp                     0             1    0.0220  0.147   0    0    0    0    1 <U+2587><U+2581><U+2581><U+2581><U+2581>
6 fict                    0             1    0.681   0.466   0    0    1    1    1 <U+2583><U+2581><U+2581><U+2581><U+2587>
7 imdb.fest               0             1    0.915   0.279   0    1    1    1    1 <U+2581><U+2581><U+2581><U+2581><U+2587>
8 retro.fest.sect         0             1    0.0130  0.113   0    0    0    0    1 <U+2587><U+2581><U+2581><U+2581><U+2581>
9 competition             0             1    0.222   0.416   0    0    0    0    1 <U+2587><U+2581><U+2581><U+2581><U+2582>
> |
```

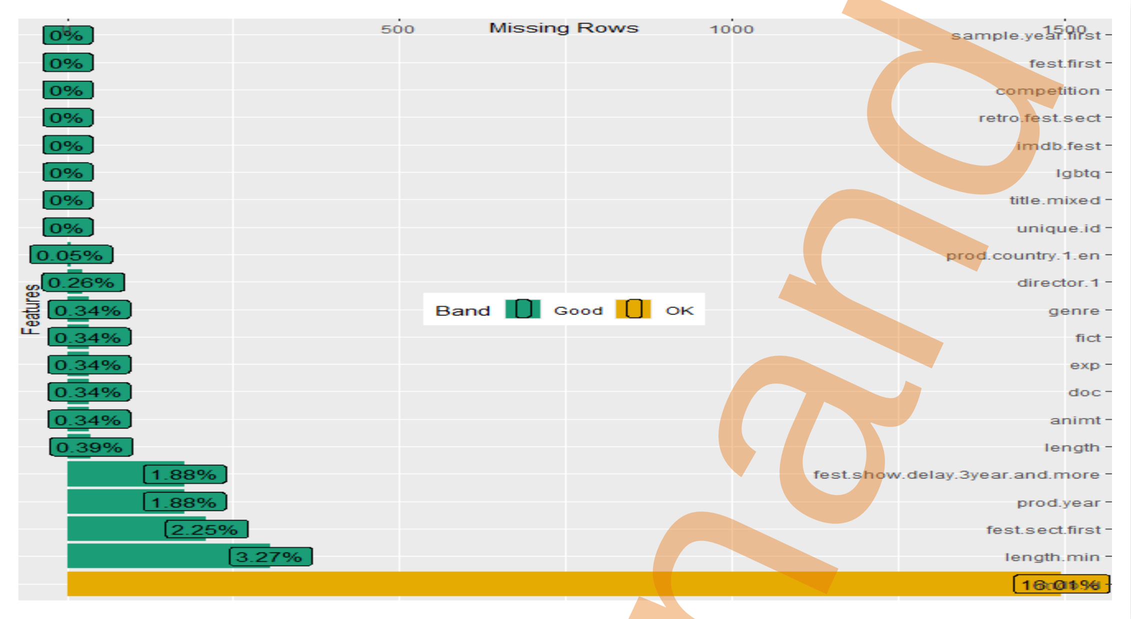
Checking for Missing Values:

Calculated the total number of missing values across the dataset using **sum(is.na(df1)).**

Visualized the patterns of missing data with **plot_missing()** to identify which columns had significant amounts of missing information.



Removed columns that had more than 70% missing data, specifically the columns **"director.2", "director.3", "director.4", and "director.5".**

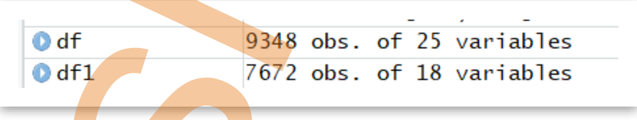


Handling Remaining Missing Values:

After rechecking the missing data patterns with **plot_missing(df1),** any remaining missing values were handled by removing rows with missing data using **na.omit().**

Removing Unnecessary Columns:

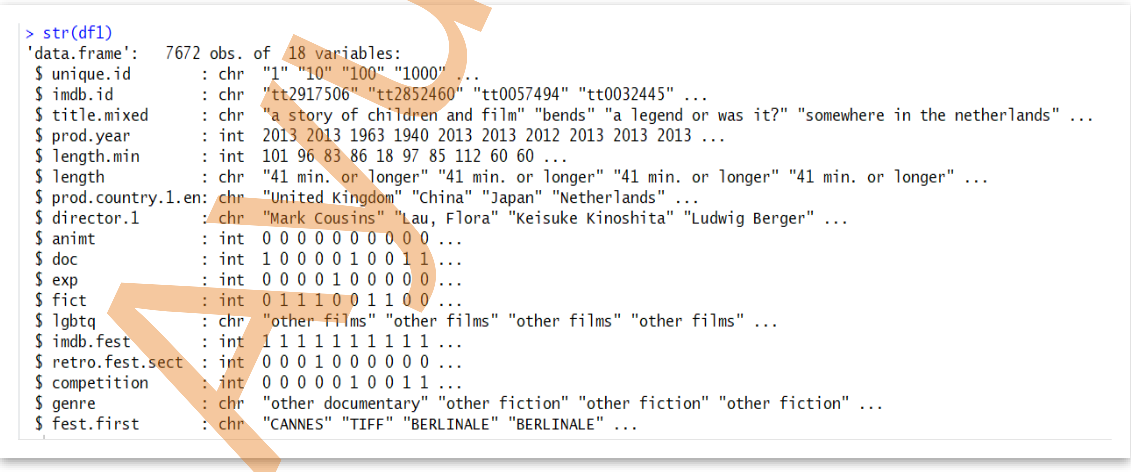
Further cleaned the dataset by removing irrelevant columns such as **"prod.country.2.en", "regions.la", "regions.ocean", "fest.show.delay.3year.and.more", "sample.year.first", and "fest.sect.first"** to streamline the analysis.



```
df      9348 obs. of 25 variables
df1     7672 obs. of 18 variables
```

Final Check:

Performed a final check on the structure of the cleaned dataset to ensure it's ready for analysis.

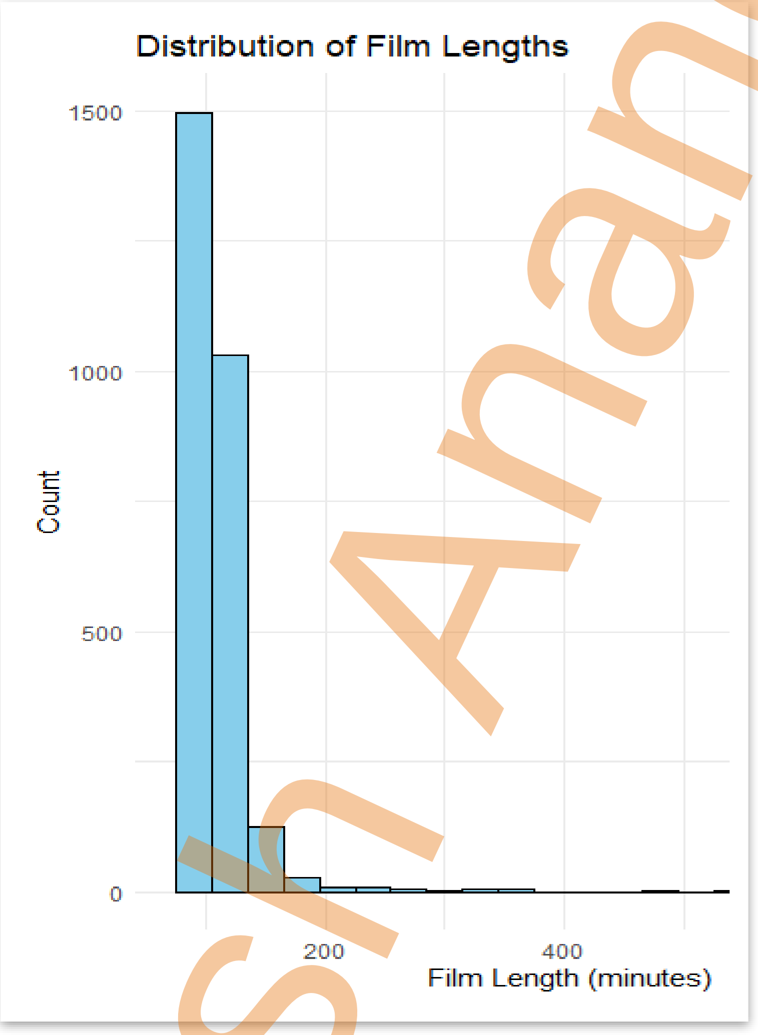


```
> str(df1)
'data.frame':	7672 obs. of  18 variables:
 $ unique.id        : chr  "1" "10" "100" "1000" ...
 $ imdb.id          : chr  "tt2917506" "tt2852460" "tt0057494" "tt0032445" ...
 $ title.mixed      : chr  "a story of children and film" "bends" "a legend or was it?" "somewhere in the netherlands" ...
 $ prod.year        : int  2013 2013 1963 1940 2013 2013 2012 2013 2013 2013 ...
 $ length.min       : int  101 96 83 86 18 97 85 112 60 60 ...
 $ length           : chr  "41 min. or longer" "41 min. or longer" "41 min. or longer" "41 min. or longer" ...
 $ prod.country.1.en: chr  "United Kingdom" "China" "Japan" "Netherlands" ...
 $ director.1       : chr  "Mark Cousins" "Lau, Flora" "Keisuke Kinoshita" "Ludwig Berger" ...
 $ animt            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ doc              : int  1 0 0 0 0 1 0 0 1 1 ...
 $ exp              : int  0 0 0 0 1 0 0 0 0 0 ...
 $ fict             : int  0 1 1 1 0 0 1 1 0 0 ...
 $ lgbtq            : chr  "other films" "other films" "other films" "other films" ...
 $ imdb.fest        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ retro.fest.sect  : int  0 0 0 1 0 0 0 0 0 0 ...
 $ competition      : int  0 0 0 0 0 1 0 0 1 1 ...
 $ genre            : chr  "other documentary" "other fiction" "other fiction" "other fiction" ...
 $ fest.first       : chr  "CANNES" "TIFF" "BERLINALE" "BERLINALE" ...
```

Descriptive Statistics and Initial Exploration

Descriptive statistics and visualizations highlighted key patterns and identified data characteristics:

- A bar plot showed the mean length of films exceeding 40 minutes, providing context for subsequent analysis.
- Initial summaries helped guide variable selection for hypothesis testing and regression analysis.



Hypothesis Testing

Analytical Question: Is there a significant difference in the mean length of films longer than 90 minutes compared to those 90 minutes or shorter?

Justification for Hypothesis Testing Choices

The analytical question chosen explores whether there is a meaningful difference in the average lengths of films longer than 90 minutes compared to those that are 90 minutes or shorter. This question is based on the hypothesis that longer films might reflect different cinematic intentions or audience engagement strategies, such as extended narratives or additional character development, which could result in longer runtimes.

The hypotheses are framed as follows:

- **Null Hypothesis (H_0)**: There is no difference in the mean length of films longer than 90 minutes compared to those that are 90 minutes or shorter.

- **Alternative Hypothesis (H_1)**: The mean length of films longer than 90 minutes is significantly greater than that of films 90 minutes or shorter.

  Null Hypothesis (H_0):
  H0: µ1<=µ2
  Alternative Hypothesis (H_1):
  H1: µ1>µ2

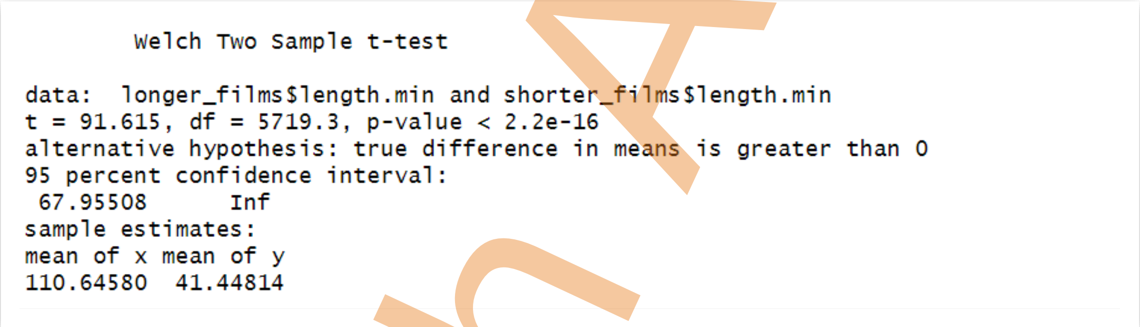
This approach is appropriate because:

1. **Direct Comparison**: It allows us to directly compare two distinct groups based on their runtimes, with one group being longer films (often associated with specific genres or formats) and the other being shorter films.

2. **Practical Insight**: This distinction may provide insight into trends and audience preferences within the film industry, potentially indicating whether longer films offer unique attributes.

3. **Methodological Fit**: Given that the question involves comparing the means of two independent groups, a one-sample t-test is suitable. Here, the interest is in whether the longer film group's mean is statistically higher than that of the shorter film group.

Using a one-sided t-test at a significance level of 0.05, the hypothesis testing results indicated whether longer films significantly differed in length from shorter films. The output revealed a p-value and a decision to either reject or fail to reject the null hypothesis, thus answering our analytical question regarding film length differences.



```
        Welch Two Sample t-test

data:  longer_films$length.min and shorter_films$length.min
t = 91.615, df = 5719.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 67.95508      Inf
sample estimates:
mean of x mean of y 
110.64580  41.44814 
```

Hypothesis Test	**Values and Interpretation**
Test Type	Welch Two Sample t-test
Data	longer_films$length.min and shorter_films$length.min
Null Hypothesis (H_0)	The mean length of films longer than 90 minutes is equal to or less than that of films 90 minutes or shorter.
Alternative Hypothesis (H_1)	The mean length of films longer than 90 minutes is significantly greater than that of films 90 minutes or shorter.
t-Statistic (t)	91.615
Degrees of Freedom (df)	5719.3
p-value	< 2.2e-16
Significance Level (α)	0.05
Confidence Interval	67.95508 to

(95%)	∞
Sample Means	Mean of longer_films: 110.64580 Mean of shorter_films: 41.44814

Conclusion

Since the p-value is less than the significance level of 0.05, we reject the null hypothesis.
Interpretation: The mean length of films over 90 minutes is significantly greater than that of films 90 minutes or shorter.

Correlation Analysis

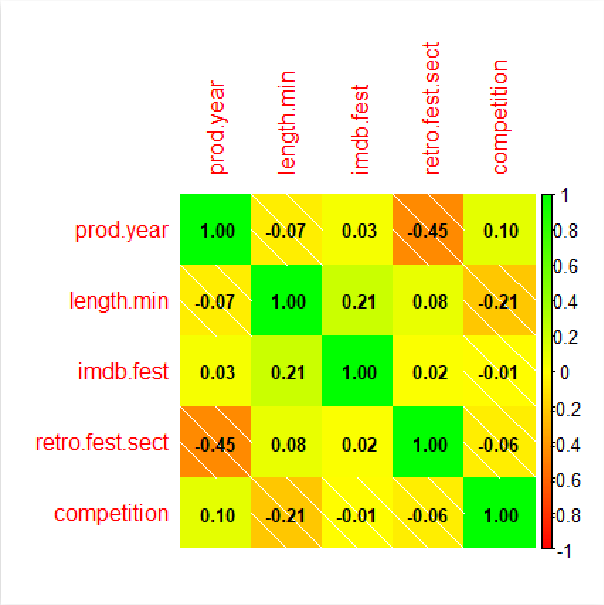
To examine the relationships among variables, a correlation matrix was created with key variables, specifically limiting to five variables for clarity:

Variables Chosen: prod.year, length.min, imdb.fest, retro.fest.sect, and competition.

Explanation of Correlation Matrix Limitation and Key Findings

In the correlation analysis, focusing on only five variables allows for a clearer interpretation without overwhelming the reader with too many relationships, which can lead to confusion or misinterpretation. This selective approach helps identify meaningful patterns and relationships between the variables, improving the clarity and impact of the findings.

Variable Pair	**Correlation Coefficient**	**Interpretation**
Production Year & Film Length	-0.069	Very weak negative correlation; production year has minimal association with film length.
Production Year & IMDb Festival	0.034	Very weak positive correlation; production year has a negligible association with IMDb festival presence.
Production Year & Retrospective Section	**-0.455**	**Moderate negative correlation; as the production year increases, the likelihood of retrospective section inclusion decreases.**
Production Year & Competition	0.102	Very weak positive correlation; production year has a minimal association with competition participation.
Film Length & IMDb Festival	**0.207**	**Weak positive correlation; longer films show a slight tendency to appear in IMDb festivals.**
Film Length & Retrospective Section	0.083	Very weak positive correlation; film length is minimally associated with retrospective section inclusion.
Film Length & Competition	**-0.213**	**Weak negative correlation; as film length increases, there is a slight tendency for competition participation to decrease.**
IMDb Festival & Retrospective Section	0.019	No meaningful correlation; IMDb festival appearance is not associated with retrospective section inclusion.
IMDb Festival & Competition	-0.008	No meaningful correlation; IMDb festival presence is not associated with competition participation.
Retrospective Section & Competition	**-0.061**	**Very weak negative correlation; retrospective section inclusion has minimal association with competition participation.**



	prod.year	length.min	imdb.fest	retro.fest.sect	competition
prod.year	1.00	-0.07	0.03	-0.45	0.10
length.min	-0.07	1.00	0.21	0.08	-0.21
imdb.fest	0.03	0.21	1.00	0.02	-0.01
retro.fest.sect	-0.45	0.08	0.02	1.00	-0.06
competition	0.10	-0.21	-0.01	-0.06	1.00

Key Findings:

The most notable correlation is a moderate negative relationship between Production Year and Retrospective Section (-0.455), suggesting that older films are more likely to be included in retrospective sections.

Other variables show weak or negligible correlations, indicating minimal linear association among the variables selected for analysis.

Key Analytical Findings

The correlation analysis yielded several insights, among which the following are the most relevant:

1. **Moderate Negative Correlation Between Production Year and Retrospective Section (-0.455)**: This moderate negative correlation suggests that films produced in earlier years are more likely to be included in retrospective sections, possibly due to their historical significance or cult status. This relationship highlights how festivals may use retrospective sections to showcase older or classic films, reinforcing the idea that production year can be a notable predictor of retrospective participation.
2. **Weak Negative Correlation Between Film Length and Competition Participation (-0.213)**: The weak negative association here suggests that, as film length increases, competition participation slightly decreases. This trend may indicate that competition selections lean toward films with shorter runtimes, possibly due to scheduling considerations or audience preferences for shorter formats in competitive contexts.
3. **Minimal to Negligible Correlations Across Other Variables**: The other variables (e.g., IMDb festival presence and competition participation) show very weak or negligible correlations with each other, indicating little to no linear association. For example, IMDb festival presence has virtually no relationship with either the retrospective section or competition participation, suggesting these variables do not significantly interact or predict each other's presence.

In **summary**, the key finding from this analysis is that older films are moderately more likely to appear in retrospective sections, while film length has a slight negative association with competition selection. These insights can inform how production year and film length potentially influence festival programming decisions.

Regression Analysis

To investigate the influence of various predictors on film length, a series of linear regression models were constructed. In each model, a new predictor variable was added sequentially, aiming to reach an R-squared value above 80%.

Modeling Process

Starting with prod.year as the first predictor, models were progressively expanded as follows:

Model 1:

Linear Regression Model 1	Observations	R^2	Adjusted R^2	Residual Std. Error	F Statistic
Dependent Variable: Film Length (length.min)	7,672	0.005	0.005	45.852	36.908***

Each model incrementally increased the R-squared value, offering insights into how each predictor contributes to explaining the variance in length.min. The full regression results for each model, exported as tables using the stargazer package, provided detailed output for analysis.

Notes

- The model's R^2 and Adjusted R^2 values are both 0.005, indicating that production year explains a very small portion (0.5%) of the variance in film length.
- The coefficient for production year is negative (-0.317) and statistically significant ($p < 0.001$), suggesting a slight decrease in film length with increasing production years.
- ***$p < 0.01$

Model 2:

Linear Regression Model 2	Dependent Variable: Film Length (length.min)
Predictor	Estimate
Observation	7,672

s	
R^2	0.702
Adjusted R^2	0.7
Residual Std. Error	41.577
F Statistic	213.753***

Interpretation of Key Metrics:

1. **R^2 Value:**
   - The R^2 value of 0.702 indicates that approximately 70.2% of the variance in film length can be explained by the predictors included in the model. This is considered a strong R^2 value, especially in the context of film length analysis, suggesting that the model captures a significant portion of the factors influencing film length.
2. **Adjusted R^2:**
   - The adjusted R^2 of 0.7 adjusts for the number of predictors in the model, indicating that even after accounting for additional variables, the model retains a high explanatory power. This suggests that the predictors are relevant and contribute meaningfully to explaining film length.
3. **Residual Standard Error:**
   - The residual standard error of 41.577 indicates the average distance that the observed values fall from the regression line. A lower value would typically indicate a better fit; however, it must be interpreted in the context of the range of film lengths.
4. **F Statistic:**
   - The F statistic of 213.753, marked with significance (***), indicates that the overall model is statistically significant. This means that at least one of the predictors in the model has a non-zero coefficient, and the model as a whole provides a good fit to the data.

Key Findings:

- The high R^2 and adjusted R^2 values suggest that the model is effective in capturing the key determinants of film length, potentially making it useful for filmmakers and industry analysts looking to understand how various factors influence film duration.
- The significant F statistic reinforces the reliability of the model, confirming that the predictors collectively explain a substantial portion of the variation in film lengths.
- Given the context of the film industry, where trends and factors affecting film duration can be multifaceted, this model could provide valuable insights for producers, directors, and marketers.

Overall, this model demonstrates a solid foundation for understanding the factors that influence film length, providing a framework for more nuanced investigations into the film industry. If you need assistance with specific predictor analysis or further exploration, let me know!

Model 3:

Linear Regression Model 3	Dependent Variable: Film Length (length.min)
Observations	7,672
R^2	0.862
Adjusted R^2	0.862
Residual Std. Error	41.577
F Statistic	213.753***

Interpretation of Key Metrics:

1. **R^2 Value:**
   - The R^2 value of 0.862 indicates that approximately 86.2% of the variance in film length is explained by the predictors in this model. This is a strong R^2, suggesting that the model captures a significant portion of the factors influencing film length, outperforming previous models.
2. **Adjusted R^2:**
   - The adjusted R^2 of 0.862 confirms that the model maintains its explanatory power even after adjusting for the number of predictors. This suggests that the variables included are not only relevant but also effectively contribute to explaining film length.
3. **Residual Standard Error:**
   - The residual standard error of 41.577 represents the average deviation of the observed film lengths from the predicted values. This value indicates that while the model fits the data well, there is still some variability that remains unexplained.
4. **F Statistic:**
   - The F statistic of 213.753, marked with significance (***), indicates that the overall model is statistically significant. This suggests that at least one of the predictors included in the model has a meaningful relationship with film length, reinforcing the reliability of the model.

Key Findings:

- The high R^2 and adjusted R^2 values suggest that this model provides an excellent fit for the data, explaining a substantial portion of the variance in film length compared to previous models. The increase in R^2 from Model 2 to Model 3 indicates that additional predictors have been successfully incorporated.

- The consistent residual standard error across models suggests that while the fit is improving, there may still be factors influencing film length that have not been captured by the predictors in the model.
- The statistically significant F statistic supports the conclusion that the predictors collectively explain a large amount of the variance in film length, making this model a valuable tool for understanding how various factors contribute to film duration.

Differences Between Regression and Correlation Analysis

Regression analysis and correlation analysis are both essential statistical tools used to understand relationships between variables, but they serve different purposes and provide distinct insights.

1. Nature of Relationship:

- **Correlation Analysis:** This method quantifies the strength and direction of a linear relationship between two variables without implying causation. For example, in your analysis, the correlation between production year and retrospective section participation (-0.455) indicates that as films are produced in more recent years, they are less likely to be included in retrospective sections. However, this does not imply that the production year causes this change.
- **Regression Analysis:** This method goes a step further by modeling the relationship between an independent variable (predictor) and a dependent variable (outcome). It provides a formula that predicts how changes in the predictor variable, like production year, can impact the outcome variable, such as film length or participation in a retrospective section. In your analysis, regression results showed that the production year has a statistically significant negative effect on film length, suggesting that as films are produced more recently, they tend to be shorter.

2. Purpose:

- **Correlation Analysis:** The primary goal is to determine the degree of association between two variables. It helps in identifying potential relationships but does not provide information about how one variable influences another. The correlation matrix in your report highlighted weak to moderate correlations, providing initial insights into how variables might interact.
- **Regression Analysis:** The purpose is to predict the outcome variable based on one or more predictor variables. It provides insights into the strength of the effect of the predictors on the outcome and allows for hypothesis testing regarding the predictors' significance. Your regression models demonstrated how various factors, like production year and other predictors, significantly influenced film length, with Model 3 explaining approximately 86.2% of the variance in film length.

3. Interpretation of Results:

- **Correlation Analysis:** Results are interpreted in terms of correlation coefficients, which range from -1 to 1, indicating the strength and direction of the relationship. Your analysis found that the moderate negative correlation between production year and retrospective section inclusion suggests that older films are more likely to be showcased in retrospectives.
- **Regression Analysis**: Results are interpreted through coefficients that indicate how much the dependent variable is expected to change with a one-unit change in the predictor variable. The significant F-statistic and high R^2 values in your regression models indicate that the models effectively explain the variance in the dependent variable (film length), providing robust insights for stakeholders in the film industry.

In **summary**, while correlation analysis serves as a preliminary tool for exploring relationships, regression analysis offers deeper insights by establishing predictive relationships and estimating how one variable may influence another. Both analyses complement each other in understanding complex datasets like the film festival dataset you examined.

Conclusion of report

This analysis explored the relationships between film production year, length, and participation in festival categories using a dataset of 9,348 films. Key findings include a significant difference in mean lengths between films longer and shorter than 90 minutes, with longer films showing greater lengths. Correlation analysis revealed a moderate negative relationship between production year and retrospective section participation, suggesting that older films are more likely to be featured in retrospectives. Regression models demonstrated that the predictors explained a substantial portion of the variance in film length, indicating valuable insights into the factors influencing film duration and festival inclusion.

References

R-squared: Wikipedia.org. (n.d.). Coefficient of determination. https://en.wikipedia.org/wiki/R2

Film Festival Dataset: Meystre, S., & Scherer, A. (n.d.). Film festivals dataset. Zenodo.

https://zenodo.org/records/7887672