**G. INSIGHTS AND REPORTING**

* Summarize the findings in a structured Format

**STATISTICAL CONCLUSION [DRAFT ALL IN ALL]**

**Chi-Square Test Results**

* **X-squared = 1793**: This is the Chi-square statistic, a measure of how much the observed data deviate from the expected data under the null hypothesis of independence.
* **df = 9**: Degrees of freedom, calculated as (number of rows - 1) \* (number of columns - 1).
* **p-value < 2.2e-16**: This extremely small p-value indicates that the observed data are very unlikely under the null hypothesis.

**Key Insights**

1. **Significant Association**:
   * The small p-value (< 0.05) strongly suggests that there is a statistically significant association between Country\_Group and Language\_Group.
   * This means that the distribution of languages is not independent of countries. Certain countries likely prefer or are associated with specific languages.
2. **Practical Interpretation**:
   * For example, countries in the "USA" group might predominantly use "English," while other countries might have more diverse language usage.
   * Further investigation is needed to understand which specific countries and languages drive this association.
3. **Explore the Data**:
   * Use proportions or visualizations to understand the relationship better. For example:

R

Copy code

prop\_table <- prop.table(table(datamovieprocessed$Country\_Group, datamovieprocessed$Language\_Group), margin = 1)

print(prop\_table)

This shows the proportion of each language within each country group.

1. **Actionable Insights**:
   * For content creators: Tailor movie offerings or marketing strategies by focusing on the dominant languages within specific countries.
   * For linguists or cultural studies: Use this insight to explore the cultural or historical ties between countries and languages.

***Next Steps***

* ***Visualization****: Create a heatmap or bar chart to visualize the relationship:*

*R*

*Copy code*

*library(ggplot2)*

*ggplot(as.data.frame(contingency\_table), aes(x = Var1, y = Var2, fill = Freq)) +*

*geom\_tile() +*

*labs(x = "Country Group", y = "Language Group", fill = "Frequency") +*

*theme\_minimal()*

* ***Focus Analysis****: Drill down into specific countries or languages driving the association.*

**Key Findings from the T-Test**

1. **T-Statistic (t = -0.046595)**:
   * The T-statistic measures the difference in means relative to the variability in the data. A value close to 0 suggests very little difference between the two groups.
2. **Degrees of Freedom (df = 50.451)**:
   * Degrees of freedom are derived based on the sample sizes and variability of the two groups. In this case, there are approximately 50 degrees of freedom.
3. **P-Value (p = 0.963)**:
   * The p-value is very high (> 0.05), indicating that there is **no statistically significant difference** between the average IMBD scores of movies in the Hindi and Mandarin language groups.
4. **Alternative Hypothesis**:
   * The test was conducted to see if the true difference in means is not equal to 0. However, the results do not support this hypothesis.
5. **Confidence Interval (-0.6344431, 0.6056682)**:
   * The 95% confidence interval includes 0, which reinforces the finding that there is no meaningful difference in the average IMBD scores between the two groups.
6. **Sample Estimates**:
   * **Mean for Hindi**: 6.774
   * **Mean for Mandarin**: 6.788
   * The averages are very close, with a difference of only 0.014, which is negligible.

**Key Insights**

1. **No Significant Difference**:
   * The T-test shows no significant difference in IMBD scores between Hindi and Mandarin movies. Both groups have similar average ratings.
2. **Practical Implication**:
   * From a movie perspective, audience ratings for movies in Hindi and Mandarin are comparable, suggesting that language may not be a strong differentiator for movie ratings.
3. **Confidence in the Result**:
   * The high p-value and the confidence interval including 0 strengthen the conclusion that the two groups are not significantly different.

***Next Steps***

* ***Explore Other Factors****: Investigate whether other variables (e.g., genre, director, or release year) may explain variations in IMBD scores.*
* ***Visualization****: Create a boxplot to visually compare the distributions of IMBD scores:*

*R*

*Copy code*

*ggplot(filtereddata, aes(x = Language, y = IMBD\_Score, fill = Language)) +*

*geom\_boxplot() +*

*labs(title = "IMBD Score Comparison by Language", x = "Language", y = "IMBD Score") +*

*theme\_minimal()*