TV Shows Project Report

Diego Sanoja

2023-02-03

Objective of the project

Analyze the data in the data_TV.csv file to understand how the TV programs have changed over the years and perform statistical analysis in the different variables to create relationships among them.

Data Used

For this project, a csv file called data_TV which contains information of the Top Rated TV Shows stored in 8 columns and 2617 rows like the release date, country and language of origin, and popularity was downloaded from a public Kaggle data set whose URL direction is https://www.kaggle.com/datasets/titassaha/top-rated-tv-shows.

The data has some limitations: its last update done to the data was done in November of 2022, it doesn't have a specified period of update.

Changelog

Version 1.0.0 (10-01-2023)

- Created a folder called TV Shows Project to store the files related to this project.

The following changes were done in Microsoft Excel.

- Used the Find and replace option to change the rows in the first_air_date column which had an NA value for an empty cell.
- Used the Find and replace option to change the rows in the origin_country column which had an character(0) string value for an empty cell.
- Saved the new workbook as a csv file with the name of data_TV_cleaned in the project folder.

Version 1.0.1 (10-01-2023)

The following changes were done using Microsoft Power BI Desktop.

- Opened the cleaned csv file in the Query Editor to:

i. Filter out the rows that were blank in the first_air_date column.

ii. Remove extra white spaces and non printable characters from the name and overview columns using the trim and clean functions.

iii. Change the original_language column format to upper case.

iv. Capitalize all the column names.

- Downloaded the following tables:

i. Table 1 from https://www.iban.com/country-codes which contains the codes of the countries in the world.

ii. Table 1 from https://en.wikipedia.org/wiki/List_of_ISO_639-1_codeswhich contains the codes for the languages in the world.

- Saved the tables and changed their names of the URL for Country_code and Language_code respectively.
- Made the following relationships between data columns using the Model View:

i. The Origin_country column of the data_TV_cleaned table to the Alpha-2 code column of the Country_code table.

ii. The Original_language column of the data_TV_cleaned table to the column 2 of the Language_code table.

- Used the Data View to make the following modifications:

i. For the Language_code table:

a. Deleted column 3 to 6.

b. Renamed columns 1 and 2 as Language_name and Language_code respectively.

ii. For the Country_code table:

a. Deleted columns 3 and 4.

b. Renamed columns 1 and 2 as Country and Alpha-2 Code respectively.

iii. For the data_TV_cleaned table:

a. Added the following calculated columns:

1. Country Full Name which contains the full name of the country where the shows were filmed. If it was filmed in more than 1 location, the cell will be blank.
2. Language Full Name which contains the full name of the language in which the program was first released.
3. Overview Length which shows the amount of characters used to give the show's premise.
4. Release Year which shows the year when the program was first released.
5. Release Season which shows the USA year season when the show was first released.

b. Added the following calculated measures:

1. Highest Popularity which computes the maximum popularity.
2. Shows Released which computes the total number of shows released.

Version 1.0.2 (11-01-2023)

- Used to the Data View to make the following modifications:

i. Replaced the names of the following columns:

a. First_air_date to First Air Date.

b. Origin_country to Origin Country.

c. Original_language to Origin Language.

d. Vote_average to Show's Vote Average.

e. Vote_count to Show's Vote Count.

ii. Added the following calculated columns:

a. Filming Locations which shows the number of countries where the program was filmed as strings.

b. Initial Filming Country which shows the name of the country where it is assumed that most of the filming took place.

ii. Added the following calculated measures:

a. % of Shows Released which computes the percentage of programs released.

b. Last Release date which computes the date of the last show released in the data set.

c. Complete Vote Average which computes the average of the program's vote average to 2 decimal places.

d. Vote Count Average which computes the average of the show's vote count to 2 decimal places.

- Used to the Model View to change the relationship between the Country_code and the data_TV_cleaned tables by switching the connection from the Origin Country to the Country Full Name column.
- Used the Report View to:

i. Create 2 pages. Their given names were Important Metrics for the first one and Graphs and Charts for the second one.

ii. Added the following visuals to the Important Metrics page:

a. A table which contains the years where the programs of the data set were released, the maximum popularity of a program, the complete average and the average vote count for a show each year.

b. Eight cards which were ordered in 2 columns, each column containing 4 cards in the following way:

1. Column 1 of cards displays the name, release date, origin country and original language of the most recent program.

2. Column 2 of cards displays the name, release date, origin country and original language of the most popular show.

c. A slicer to filter the visuals of the page based on the first air dates of the shows.

iii. Added the following visuals to the Graphs and Charts page:

a. A line chart that displays the number of shows in the y-axis and the year of release in the x-axis along with the percentage of shows released each year.

b. A line chart that shows the number of programs in the y-axis and the year of release in the x-axis along with 4 lines to display the number of shows released every season of every year.

c. A map which uses bubbles to display the country where a program was filmed/ produced, the full name of the country, the number of shows produced and the percentage of releases.

d. A donut chart to show the percentage of programs based on the number of countries used to film it.

e. A vertical bar chart which displays the number of shows produced in the 6 most used languages and their percentage of the total set.

- Exported the data in the data_TV_cleaned table to a excel workbook.
- Created a word document called DAX Codes Used which contains the DAX codes used to create the calculated columns and the measures in Power BI along with the name of their respective measure or calculated column in the same order in which they were created.

Version 1.1.0 (17-01-2023)

- Removed the following columns from the data_TV_cleaned file:

i. First Air Date.

ii. Origin Country.

iii. Original Language.

iv. Overview.

v. Release Year.

- Changed the name of the following columns:

i. Name to NAME.

ii. Popularity to POPULARITY.

iii. Show's Vote Average to SHOWVOTEAVERAGE.

iv. Show's Vote Count to SHOWVOTECOUNT.

v. Country Full Name to COUNTRY.

vi. Language Full Name to LANGUAGE.

vii. Overview Length to OVERVIEWLEN.

viii. Release Season to RELSEASON.

ix. Filming Locations to FILMLOC.

x. Initial Filming Country to FIRSTFILMLOC.

- Saved the new Excel workbook in the project folder with the name of data_TV_cleaned_SAS.
- Uploaded the data_TV_cleaned_SAS workbook to SAS Studio.

Version 1.2.0 (17-01-2023)

- Created a new SAS file called TV SHOWS PROJECT in SAS Studio.
- Added a new data set called STAT which contains the data_TV_cleaned_SAS workbook data and a categorical variable called ENGLISH which equals 1 if the language is English and 0 otherwise.
- Removed the labels of the variables in the STAT data set using the procedure DATASETS.
- Wrote the following procedures that performed the following tasks on the STAT data set:

i. FREQ which created a tables which show the proportions of the following variables:

a. LANGUAGE.

b. FILMLOC.

c. RELSEASON.

ii. UNIVARIATE to compute descriptive statistics of the POPULARITY variable and check if it came from a normal population.

iii. CORR which computed descriptive statistics and the correlations between the following variables:

a. SHOWVOTECOUNT.

b. ENGLISH.

c. OVERVIEWLEN.

iv. REG which performed linear regression in the following variables:

a. POPULARITY on SHOWVOTECOUNT.

b. GLM which performed linear regression in the following variables:

c. POPULARITY on SHOWVOTECOUNT and OVERVIEWLEN.

d. POPULARITY on SHOWVOTECOUNT and ENGLISH.

vi. Repeated the second procedure of part v but filtered the data set by removing outliers which were considered as the points where popularity was higher than 1000 and removed the ENGLISH variable from the procedure.

vii. FREQ which computed a 2-way frequency table of the FILMLOC and RELSEASON variables.

- Added a new data set called STAT2 which has the same information as STAT plus the following variables:

i. LPOPULARITY which computes the natural logarithm of the POPULARITY.

ii. A categorical variable called GROUPS which concatenated the values of the FILMLOC and RELSEASON variables.

- Wrote the following procedures that performed the following tasks on the STAT2 data set:

i. UNIVARIATE in the LPOPULARITY variable to compute descriptive statistics like the mean and quantiles of the variable to understand its distribution.

ii. GLM which performed linear regression in the following variables:

a. LPOPULARITY on SHOWVOTECOUNT and ENGLISH.

iii. REG which performed the following tasks:

a. Linear regression of LPOPULARITY on SHOWVOTECOUNT.

b. Created a data set named RESIDUALS1 which contains the residuals and predicted values of the dependent variable (LPOPULARITY) after performing the regression.

iv. UNIVARIATE to perform a normal test in the values of the RESIDUALS1 data set.

v. GLM which performed one-way ANOVA (Analysis of Variance) in the SHOWVOTEAVERAGE using the categorical variables FILMLOC, RELSEASON and GROUPS. Furthermore, data sets which contain the residuals for those processes were also created with the names of RESIDUALS2, RESIDUALS3, and RESIDUALS4 respectively.

- Performed the following statistical tests using a significance level of 0.05 on the following procedures after writing them:

i. Linear Relationship test to check for a linear relationship between the independent and dependent variables on each REG and GLM procedure.

ii. A test to check if the coefficients of each regression parameters were different than 0 for each REG and GLM procedures that were used for simple linear regression.

iii. A test to check for interactions between regression parameters in the GLM procedures that were used for simple linear regression with more than 1 independent variable.

iv. The Kolmogorov-Smirnov test for normality for the 4 data sets that contained the residuals of the REG and GLM procedures.

v. The equality of means test for the ANOVA done to the FILMLOC, RELSEASON and GROUPS variables using the GLM procedures.

vi. The Bartlett's test for homogeneity of variance done to each ANOVA analysis.

Version 1.2.1 (24-01-2023)

- Moved the procedures, relevant images of the outputs, and the test results to a Word document called SAS Codes, Tables and Results. Then saved the document in the project folder.
- Generated and saved a pdf using the Word file with the same name in the project folder.

Analysis Summary

For the analysis, 3 different programs were used. The first one was Excel where the data in the data_TV file was cleaned by removing rows with inconsistent data. The second one was Microsoft Power BI Desktop where the combined data of data_TV and 2 tables downloaded from the internet was cleaned, transformed, visualized in 2 interactive dashboards and exported to another Excel workbook. The third tool was the SAS Studio page where statistical analysis was performed in the data of the exported csv file by performing simple linear regression and ANOVA analysis in some variables of the data using the SAS programming language.

Analysis Summary in Power BI

For the analysis phase using Power BI, each image in each interactive dashboard was analyzed. The first dashboard analyzed was Important Metrics and the second was Graphs and tables.

Important Metrics

This dashboard contains 10 visuals, each one which helps to uncover relevant insights from the data.

The first visual analyzed in this dashboard was a table which has the release year of the TV shows, the number of the biggest popularity and the averages of the show's vote average and the vote count of each year.

By sorting the columns of the table, the following insights were discovered:

i. The release years of the programs in the data set start from 1951 and end in 2022. Furthermore, there were years where no shows were released. Those years were from 1952 to 1954 and 1970.

ii. The air year of the most popular program was 2022 with a popularity of 6684.61. Furthermore, the rest of the shows in the data set have popularity lower than 1600.

iii. The year with the greatest and lowest complete vote averages were 1959 and 1963 with an average of 8.05 and 7.13 respectively.

iv. The year with the greatest and lowest vote count averages were 1989 and 1978 with an average of 1134.56 and 135.14 respectively.

The next visuals analyzed were the 8 cards which were arranged in 2 columns of 4 cards per column. Each card was filtered and arranged based on a specific measure of a show. The cards were arranged in the following way: name, first air date, country of origin, and original language of the show.

After analyzing the information displayed by the columns, it was concluded that:

i. Dahmer-Monster: The Jeffrey Dahmer Story was the most recent program in the data set which was released on September 21, 2022. Furthermore, it's country of origin and original language are the USA and English respectively.

ii. House of Dragons was the most popular program in the data set which was released on August 21, 2022. Furthermore, it's country of origin and original language are the USA and English respectively.

Furthermore, the card columns can be filtered by year by selecting the Release Year column of the table to find the same information of the shows for the selected years.

The last visual in the dashboard is a slicer which can be used to filter both dashboards based on the first air date of the shows.

Graphs and Charts

This dashboard contains 5 visuals, each one which helps to uncover relevant insights in the data.

The first visual analyzed in this dashboard was the line graph located at the top left of the page. This graph displays the number of shows released per year and also it can also show the percentage of the programs released in that year by touching the point of the year of interest.

By examining this visual, the following conclusions were reached:

i. Most of the TV programs in the data set were released after 2000 with 2020 being the year where the greatest number of released (263 which is equivalent 10.07% of the observations) occurred.

ii. The number of shows decreased significantly during the year of 2021 and 2022. A reason for this decrease most likely was the COVID pandemic since there were many programs that were either cancelled or had their production delayed.

Furthermore, using the information displayed by the first card column in the Important Metrics dashboard, it was also observed that the data set only covered until the third quarter of the 2022 which could be another reason for why this year there were less releases than the previous.

The second visual examined was the second line graph located at the top right of the page. This graph displays the number of programs released both per year and season in the USA.

After studying the line chart closely, the following insights were discovered:

i. Through most of the years, the season with more program releases was Autumn.

ii. Through most of the years, at least one TV series was released during each season, especially after 1995.

iii. The year with the greatest number of series aired for spring, summer, and winter was 2020 with 69, 55, and 79 shows respectively. For the autumn season, the year with more shows aired was 2019 with a total of 67 series.

To continue with the analysis, the map with bubbles located at the bottom left was studied. This map was designed to display the number of shows that were filmed in each country. In the cases were there was more than one country used for the filming, the country where it is assumed that most of the filming occurred was used. The bigger the bubble, the greater the number of shows filmed in the country. If a bubble is selected, it will show the complete name of the country, the number and percentage of shows filmed on it.

The study of the map provided the following insights:

i. The country with the greatest number of shows filmed was USA (the United States of America) with a total of 1407 shows filmed on it which is equivalent to 53.89% of the total data.

ii. The country with the second greatest number of shows filmed was Japan with a total of 396 shows filmed on it which is equivalent to 15.17% of the total data.

iii. At least one country of each continent in the world was used as a filming location.

The second to last visual created for the analysis was a pie chart which shows the proportion of TV programs by filming locations.

After studying the pie chart closely, the following conclusions were reached:

i. Most of the shows (2490 which is equivalent to 95.37% of the data set) were filmed in 1 country.

ii. Out of the remaining 4.63% of the TV series in the set, 3.37% were filmed in 2 different countries.

The last image studied in the dashboard was a vertical bar chart which displays the number of shows whose original languages were the six most used in the data set which were: English, Japanese, Spanish Castilian, Korean, Portuguese, and French. Furthermore, if the bars are touched is it possible to see the percentage of shows released on those languages.

By examining the bar chart carefully, the following insights were discovered:

i. English was the most used original language of the programs with a total of 1682 which is equivalent to 64.42%. Furthermore it is the only language used to produce more than 1000 programs.

ii. Japanese, Spanish Castilian, and Korean were the second, third, and fourth most used languages to produce the programs. Besides English, those languages were the only ones which were used to film more than 100 shows.

iii. Portuguese, French and each of the remaining languages were used in the production of less than 25 shows.

iv. By using both this bar chart and the map, it is possible to conclude the following:

a. English was used in each continent to film at least 1 program.

b. Most of the shows whose main language was Japanese, were filmed only in Japan.

c, The shows released in Spanish were filmed Central or South America, and Europe.

Analysis Summary in SAS Studio

For the analysis phase using SAS studio, the information in the data_TV_cleaned_SAS csv file was uploaded and then analyzed using statistical methods like hypothesis testing, linear regression and ANOVA (Analysis of Variance). This phase was divided in 3 sections:

i. Basic Overview which was focused on gathering a basic idea of the structure in the data columns that were going to be studied using regression and ANOVA and creating new variables like the ENGLISH categorical variable.

ii. Linear Regression which was focused on performing linear regression on the dependent variable POPULARITY and the independent variables SHOWVOTECOUNT, OVERVIEWLEN and the new categorical variable ENGLISH.

iii. ANOVA which focused on performing statistical tests to compare the means of the SHOWVOTEAVERAGE for the categorical variables RELSEASON and FILMLOC.

Basic Overview

The information in the data_TV_cleaned_SAS csv file was stored in a data set called STAT. Furthermore, a categorical variable called ENGLISH was added to it which equals 1 if the original language of the show was English and 0 otherwise.

The first variables analyzed were the categorical variables LANGUAGE, RELSEASON and FILMLOC using the FREQ procedure by creating frequency tables for each of them whose rows were ordered in alphabetical order to understand their distribution. Each table has 5 columns, which were arranged in the following order:

i. Categorical Variable.

ii. Frequency (number of appearances in the set).

iii. Percent.

iv. Cumulative frequency (the frequency number of the previous rows was added to the next one as the table descended).

v. Cumulative percent.

Those frequency tables provided the following insights of the categorical variables:

i. There were 23 different original languages for the TV shows in the data set where the most used was English.

ii. Most of the programs were filmed in a single country. Furthermore, only the location of 21 of them (Which is equivalent to 0.8% of the observations) was not revealed.

iii. Autumn was the season of the year where most programs had their first air date. Furthermore, the difference in the number of shows released in the other seasons was small.

The next variable examined was POPULARITY which was later used for simple linear regression. For this column, the UNIVARIATE procedure was used to calculate important measures like the mean (the average of a set of values), standard deviation (positive square root of the variance), and the variance (a measure

of how spread are the values in the set), and to produce tables which display some relevant quantiles (cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way) and the 5 lowest and highest observations of this popularity in the set.

After examining the information provided by the procedure used above, the following conclusions were reached:

i. The mean equals 59.9205936.

ii. The standard deviation and variance equal 222.651355 and 49573.6259 respectively.

iii. The 0.75 quantile (75 percentile) equals 49.891 which is lower than the mean. Therefore, the data is skewed to the left.

iv. The 5 lowest observations have values lower than 4 and the 5 highest ones are higher than 1500. Those 5 extreme high observation are likely to be outliers and part of the reason of why the mean is higher than the median (is the value separating the higher half from the lower half of a data sample).

The last step of this section was to perform a correlation test between the variables that would be used as the independent variables for the linear regression process. For this the CORR procedure was used to compute a symmetrical table with the sample correlations between the each variable.

By using the table and hypothesis testing with a 0.05 level of significance, it was discovered that the correlation between the OVERVIEWLEN and SHOWVOTECOUNT variables with the ENGLISH one was not 0 while the first 2 mentioned are not correlated with each other. Despite the fact that there is correlation between some variables, the sample correlation is small and should not affect significantly the linear regression procedures.

Linear Regression

For this section multiple procedures were used to perform linear regression (is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables) on the POPULARITY variables with the other 3 independent variables. The procedures used were:

i. REG to perform regression with only 1 independent variable.

ii. GLM to REG to perform regression with more than 1 independent variable. Furthermore this procedure was used to test if there were interactions between the independent variables (this is equivalent to say that their cross products were important or not in the regression).

For each procedure used to perform linear regressions, tests were conducted to analyze the following characteristics of each regression model.

i. Fitness of the model.

ii. Values of regression parameter coefficients.

iii. Interaction between the independent variables (only done for the GLM procedures).

The following linear models were studied using regression:

i. POPULARITY on SHOWVOTECOUNT.

ii. POPULARITY on SHOWVOTECOUNT and OVERVIEWLEN.

iii. POPULARITY on SHOWVOTECOUNT and ENGLISH.

After performing the tests mentioned above on each linear model the following conclusions were reached:

i. For each model, the R-square value which is a measure of how good is the linear relation between the variables was lower than 0.1 which showed that the independent variables had a small relation with the dependent one.

ii. Adding the OVERVIEWLEN variable to the model when SHOWVOTECOUNT was already in it didn't increase the R-square value significantly which implies that OVERVIEWLEN didn't improve the relationship with the dependent variable since the coefficient of its regression parameter equals 0 by the second test. Furthermore, there was no interaction between both independent variables.

iii. Adding the interaction term between SHOWVOTECOUNT and the ENGLISH variables to the model when SHOWVOTECOUNT was already in it increased the R-square value slightly which implies that this interaction improves the relationship with the dependent variable by a small amount. However, the ENGLISH variable doesn't contribute much to the model since the coefficient of its regression parameter equals 0 by the second test.

Given the results of the procedures above and the analysis done in the Basic Overview section, it was concluded that maybe the outliers of the Popularity variable were affecting the regression analysis. Therefore, the last linear regression procedure was repeated but the outliers were removed which were observations of the dependent variable with a value higher than 1000. By examining the new results and removing 9 outliers from the sample, the following conclusions were reached:

i. The linear relationship improved since the R-square value increased to 0.365412.

ii. The coefficient of each regression parameter was different than 0. Therefore, the ENGLISH variable now was relevant to the model as well as the interaction term between the independent variables.

To conclude with the linear regression analysis, a new data set called STAT2 was created, this set had 2 additional variables. The first was named LPOPULARITY which is equal to the natural logarithm of POPULARITY.

By using the UNIVARIATE procedure to examine the distribution and structure of this new variable and comparing it to the original, the following insights were found:

i. The mean, standard deviation and variance equal 3.440949, 0.91716 and 0.84119 respectively which were lowers than the one POPULARITY, specially for the standard deviation and variance.

ii. The mean was higher than the median but the difference between them was smaller than before.

iii. The 5 lowest observations have values lower than 1.5 and the 5 highest ones are higher than 7 which are lower than before, specially for the highest observations.

The second to last step was to perform the linear regression on this new variable with the SHOWVOTECOUNT and ENGLISH variables using the GLM procedure. The results of this procedure provided the following insights:

i. There is a linear relationship between the dependent and independent variables. Furthermore the relationship is stronger using natural logarithm of POPULARITY instead since its R-square value equals 0.278721 without having to remove any observations.

ii. The ENGLISH variable and the interaction term between the variables were not relevant any more for the model since the coefficient of its regression parameters were 0.

The final part examination of this section was to confirm if this model (the one LPOPULARITY with SHOWVOTECOUNT) satisfied the assumptions of the normal linear regression model which were the following:

i. The errors are independents but that is always assumed.

ii. The errors have a constant variance which is checked by confirming if the residuals have a constant variance since the residuals are estimates of the error terms.

By looking at the graphs of the residuals against the predicted values or the SHOWVOTECOUNT variable it is shown that the residuals don't have a constant variable since the residuals tend to have larger negative values when the predicted values or the SHOWVOTECOUNT are at high values.

iii. The errors come from a normal distribution which is checked by confirming that the residuals come from a normal distribution.

By creating a new data set called RESIDUALS1 using the residual values of the last procedure and running the Kolmogorov-Smirnov test for normality on those observations, it was concluded that the residuals don't come from a normal distribution.

Since the 2 last assumptions are violated, the normal linear regression model is not a good fit for the data in the model.

ANOVA

The ANOVA (Analysis of Variance) is a statistical method that separates observed variance data into different components to use for additional tests. For this section of the analysis, 3 one-way ANOVA models (models with 1 dependent and 1 independent variable) were built and analyzed. The dependent variable was always SHOWVOTEAVERAGE and the independent variables for the 3 models were FILMLOC, RELSEASON, and GROUPS (the second additional variable of the set STAT 2 which is created by joining the different classes of the FILMLOC and RELSEASON variables). The objective of each model was to determine if there was any difference in the means of the SHOWCOUNTAVERAGE for each group.

After performing this statistical method using the GLM procedure, 3 assumptions had to be satisfied to confirm of the data was also a good fit for the model. Those assumptions are the following:

i. The errors are independents but that is always assumed.

ii. The errors of each group have a constant variance which was checked by performing the Bartlett test for homogeneity of variance in the residuals of each group since the residuals are estimate points of the error terms.

iii. The errors come from a normal distribution which was checked by confirming if the residuals came from a normal distribution.

To perform the tests for the last 2 assumptions, additional data sets were computed from the results of the GLM procedures which contained the residuals generated by each procedure. The Bartlett and the normality tests were performed on those sets to check the assumptions.

The first 2 one-way models analyzed were the ones where the independent variables were FILMLOC and RELSEASON. The ANOVA models produced using those categorical variables and the tests done to them helped to reach the following conclusions:

i. The mean values of the SHOWVOTEAVERAGE were equal for both models meaning that each category of the FILMLOC and RELSEASON variables have similar mean values of the dependent variable.

ii. The Bartlett tests done to the RESIDUALS2 (set with residuals of FILMLOC model) and the RESIDUALS3 (set with residuals of RELSEASON model) showed that the residuals of each group for each model had equal variances.

iii. The Kolmogorov-Smirnov test for normality done to the RESIDUALS2 and RESIDUALS3 data sets showed that the residuals for both models didn't come from a normal distribution which means that the third assumption was violated for both models. However, the sample size is large, and ANOVA is robust to the normality assumption meaning the model tolerates violations to this assumption. Therefore, the ANOVA model is a good fit for the data in both cases.

The original objective was to also perform a two-way ANOVA model with the independent variables of the previous models. However, it was necessary to check first if there were observations in each sub group. To confirm this, a two-way frequency table was created using the FREQ procedure and the FILMLOC and RELSEASON variables.

By carefully examining the two-way frequency table, it was discovered there weren't TV shows that were filmed in 3 countries and got first aired in spring nor programs that were filmed in 4 or more countries and

got first aired in summer. Therefore, performing a two-way ANOVA analysis was not possible. However, a one-way model could be built by using the GROUPS variable.

The ANOVA model produced using the GROUPS categorical variable and the tests done to them helped to discover the following insights:

i. The mean values of the SHOWVOTEAVERAGE were equal for each group of the categorical variable.

ii. The Bartlett test done to the RESIDUALS4 (set with residuals of GROUPS model) showed that the residuals of each group had equal variances.

iii. The Kolmogorov-Smirnov test for normality done to the RESIDUALS4 data set showed that the residuals of the model didn't come from a normal distribution which means that the third assumption was violated for both models. However, the sample size is large, and ANOVA is robust to the normality assumption. Therefore, the ANOVA model is a good fit for the data in this case.

Most Relevant Conclusions

i. House of Dragons was the most popular program in the data set which was released on August 21, 2022. Furthermore, it's country of origin and original language are the USA and English respectively.

ii. The number of TV shows had increased significantly over the last years (there was a decrease starting in 2020 most likely due to the COVID pandemic).

iii. Most of the shows were filmed in USA and the most used original language was English.

iv. Through most of the years, the season of the year with more program releases was Autumn.

v. The natural logarithm of the popularity of a show has a linear relationship of small strength with the show's vote count and no relationship with the length of the show's overview.

vi. The mean of the show's vote average is equal for each program based on the number of countries used to film them, their air seasons, and the combination of the previous factors.

Recommendations to improve the analysis

i. Gather further information of each program like number of seasons released until today, average length of the episodes, gender, and the name of the director.

ii. Increase the number of observations to find more insights and reach stronger and better informed conclusions.

iii. Use different functions to transform the data and develop new relationships between the variables.