```
FILENAME REFFILE '/home/u60677466/sasuser.v94/data_TV_cleaned_SAS.xlsx';
PROC IMPORT DATAFILE=REFFILE
        DBMS=XLSX
        OUT=WORK.IMPORT1;
        GETNAMES=YES;
RUN;
```

This proc imports the excel workbook with the observations and saves the data in a new data set called IMPORT1 located in the WORK library.

```
PROC DATASETS LIBRARY = WORK;
   MODIFY IMPORT1;
   ATTRIB _ALL_ LABEL = " ";
RUN;
```

This proc removes all the labels from the imported excel file.

```
PROC CONTENTS DATA=WORK.IMPORT1;
RUN;
```

Basic Overview

```
DATA STAT;
SET WORK.IMPORT1;
IF LANGUAGE = 'English' THEN ENGLISH = 1;
ELSE IF LANGUAGE ^= 'English'  THEN ENGLISH = 0;
RUN;
```

This proc creates a data set called STAT by using the data in WORK.IMPORT1 and then adds the conditional variable ENGLISH.

```
PROC FREQ DATA=STAT;
TITLE 'Frequency Tables of Categorical Variables';
TABLE LANGUAGE FILMLOC RELSEASON;
RUN;
```

This proc creates frequency tables of the specified categorical variables, the tables are sorted in alphabetical order.

LANGUAGE				
LANGUAGE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Arabic	2	0.08	2	0.08
Catalan, Valencian	2	0.08	4	0.15
Chinese	10	0.38	14	0.54
Danish	6	0.23	20	0.77
Dutch, Flemish	1	0.04	21	0.80
English	1682	64.42	1703	65.22
French	20	0.77	1723	65.99
German	13	0.50	1736	66.49
Hebrew	2	0.08	1738	66.56
Hindi	4	0.15	1742	66.72
Icelandic	2	0.08	1744	66.79
Italian	11	0.42	1755	67.22
Japanese	397	15.20	2152	82.42
Korean	99	3.79	2251	86.21
Norwegian	5	0.19	2256	86.40
Polish	2	0.08	2258	86.48
Portuguese	22	0.84	2280	87.32
Russian	5	0.19	2285	87.51
Spanish, Castilian	298	11.41	2583	98.93
Swedish	6	0.23	2589	99.16
Tagalog	1	0.04	2590	99.20
Thai	2	0.08	2592	99.27

LANGUAGE				
LANGUAGE	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Turkish	19	0.73	2611	100.00

FILMLOC				
FILMLOC	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1	2490	95.37	2490	95.37
2	88	3.37	2578	98.74
3	5	0.19	2583	98.93
4 or more	7	0.27	2590	99.20
Not given	21	0.80	2611	100.00

RELSEASON				
RELSEASON	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Autumn	936	35.85	936	35.85
Spring	573	21.95	1509	57.79
Summer	525	20.11	2034	77.90
Winter	577	22.10	2611	100.00

```
PROC UNIVARIATE DATA=STAT;

TITLE 'Descriptive Statistics of POPULARITY’;

VAR POPULARITY;

RUN;
```

Basic Statistical Measures			
Location		Variability	
Mean	59.92059	Std Deviation	222.65135
Median	27.51800	Variance	49574
Mode	17.84900	Range	6684
		Interquartile Range	33.28900

Quantiles (Definition 5)	
Level	Quantile
100% Max	6684.611
99%	550.034
95%	169.907
90%	104.810
75% Q3	49.891
50% Median	27.518
25% Q1	16.602
10%	11.497
5%	9.242
1%	6.326
0% Min	0.866

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.866	320	1512.00	824

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1.184	109	2493.03	1203
1.948	1076	4787.46	340
2.252	779	5865.90	956
3.988	1138	6684.61	859

For the following proc, a test will be performed using a significance level of 0.05.

```
PROC CORR DATA=STAT;
TITLE 'Correlation Table Between Independent Variables';
VAR SHOWVOTECOUNT OVERVIEWLEN ENGLISH;
RUN;
```

This proc computes descriptive statistics for each of the specified variables and a table which displays a symmetric matrix with the sample correlation coefficient between the variables.

Pearson Correlation Coefficients, N = 2611 Prob > |r| under H0: Rho=0			
	SHOWVOTECOUNT	OVERVIEWLEN	ENGLISH
SHOWVOTECOUNT	1.00000	-0.01392 0.4771	0.07538 0.0001
OVERVIEWLEN	-0.01392 0.4771	1.00000	-0.14551 <.0001
ENGLISH	0.07538 0.0001	-0.14551 <.0001	1.00000

Ho (Null Hypothesis): the correlation between both variables equals 0.

Ha (Alternative Hypothesis): the correlation between both variables is different than 0.

alpha = 0.05:

Decision Rule: Reject Ho if alpha is higher than the p-value.

The p-value is lower than alpha for columns 1 and 2 in the last row but not for the first column in the second row. Therefore, reject Ho for the correlations with the ENGLISH variable.

Conclusion: the ENGLISH variable is correlated with the other 2 variables, but the other variables are not correlated with each other.

By looking at the correlation table, it is concluded that each of the variables have small correlation.

Those variables will be used in the next procs for linear regression.

Linear Regression

For the following 10 procs, 2 or 3 tests will be performed using a significance level of 0.05.

Test 1: Fitness of the model.

Ho (Null Hypothesis): there is no linear relationship between the variables.

Ha (Alternative Hypothesis): there is a linear relationship between the variables.

Test 2: value of regression coefficients.

Ho (Null Hypothesis): the regression coefficient is equal to 0.

Ha (Alternative Hypothesis): the regression coefficient is different to 0.

Test 3: interaction between the independent variables:

Ho (Null Hypothesis): there is no interaction between the independent variables.

Ha (Alternative Hypothesis): there is an interaction between the independent variables.

Note: the Test 3 will only be done for the procs where linear regression with more than 1 independent variable is performed.

alpha = 0.05:

Decision Rule: Reject Ho if alpha is higher than the p-value.

```
PROC REG DATA=STAT;
TITLE "Regression of Popularity with Show's Vote Count";
MODEL POPULARITY = SHOWVOTECOUNT;
RUN;
```

To solve the Test 1, use the p-value in the Analysis of Variance Table.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	10559732	10559732	231.85	<.0001
Error	2609	118827432	45545		
Corrected Total	2610	129387164			

The p-value is less than 0.0001. Therefore, reject Ho.

Conclusion: there is a linear relation between Popularity and Show's Vote Count.

To solve the test 2, use the p-value in the Parameter Estimates Table.

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	28.44918	4.65999	6.10	<.0001
SHOWVOTECOUNT	1	0.05195	0.00341	15.23	<.0001

The p-value is less than 0.0001 for both parameters. Therefore, reject Ho for both.

Conclusion: each coefficient of the regression parameters is different to 0.

However, the R-Square value (Sum of Squares of Model divided by the one of the Corrected Total) which is a measure of how good the linear relation between the variables is equals 0.0816 which indicates the variance is not well explained by the regression.

```
PROC GLM DATA=STAT;
TITLE "Regression of Popularity with Show's Vote Count and Overview Length";
MODEL POPULARITY = SHOWVOTECOUNT | OVERVIEWLEN;
RUN;
```

Test 1:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	10565701.1	3521900.4	77.27	<.0001
Error	2607	118821462.5	45577.9		
Corrected Total	2610	129387163.5			

The p-value is less than 0.0001. Therefore, reject Ho.

Conclusion: there is a linear relation between Popularity and both Show's Vote Count the Overview Length.

Test 2:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	28.74733271	8.21349829	3.50	0.0005
SHOWVOTECOUNT	0.04978558	0.00719112	6.92	<.0001
OVERVIEWLEN	-0.00095887	0.02239170	-0.04	0.9658
SHOWVOTEC*OVERVIEWLE	0.00000729	0.00002129	0.34	0.7321

The p-value is less than alpha for the regression coefficient of the intercept and the regression parameter of the Show's Vote Count variable while for the rest of the coefficients, alpha is lower than the p-value. Therefore, reject Ho for every regression coefficient of the Overview Length and the interaction variable but don't for the other regression parameters.

Conclusion: only the regression coefficients for the intercept and the Show's Vote Count variables are different than 0.

To solve Test 3, either use the results of Test 2 or the p-value of the table with the TYPE III SS column. Both tables reach the same conclusion. Therefore, reject Ho.

Conclusion: there is no interaction between the Overview length and the Show's Vote Count Variables.

However, the R-Square value now is 0.08166 which indicates that adding the Overview Length variable doesn't significantly improve the relationship with the dependent variable, a fact that is further supported by the results of Test 2.

```
PROC GLM DATA=STAT;
TITLE "Regression of Popularity with Show's Vote Count and the English Variable";
MODEL POPULARITY = SHOWVOTECOUNT | ENGLISH;
RUN;
```

Test 1:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	11044892.5	3681630.8	81.10	<.0001
Error	2607	118342271.1	45394.0		
Corrected Total	2610	129387163.5			

The p-value is less than 0.0001. Therefore, reject Ho.

Conclusion: there is a linear relationship between the Popularity and the independent variables.

Test 2:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	30.65376501	7.82891301	3.92	<.0001
SHOWVOTECOUNT	0.03220827	0.00731937	4.40	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
ENGLISH	-1.58744658	9.75006700	-0.16	0.8707
SHOWVOTECOUN*ENGLISH	0.02480486	0.00827574	3.00	0.0027

The p-value is less than alpha for every regression parameter except for the one of the English categorical variable. Therefore, reject Ho only for the regression parameter of the English variable.

Conclusion: all the coefficients of the regression parameters are different than 0 except for the one of the English categorical variable.

Test 3:

Using the results of Test 2, reject Ho.

Conclusion: there is an interaction between the Show's Vote Count and the English Categorical variable.

Furthermore, the R-Square value is 0.085363 which is higher than the one of the first regression procedure which means that adding the interaction term English variable in the regression improves the relationship.

PROC GLM DATA=STAT;

TITLE "Regression of Popularity Without Outliers";

MODEL POPULARITY = SHOWVOTECOUNT | ENGLISH;

WHERE POPULARITY LE 1000;

RUN;

Test 1:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5455181.41	1818393.80	498.66	<.0001
Error	2598	9473678.82	3646.53		
Corrected Total	2601	14928860.23			

The p-value is less than 0.0001. Therefore, reject Ho.

Conclusion: there is a linear relationship between the Popularity and the independent variables.

Test 2:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	30.65376501	2.21892074	13.81	<.0001
SHOWVOTECOUNT	0.03220827	0.00207450	15.53	<.0001
ENGLISH	-6.49150060	2.77933213	-2.34	0.0196
SHOWVOTECOUN*ENGLISH	0.01092303	0.00240678	4.54	<.0001

The p-value is less than alpha for every regression parameter. Therefore, reject Ho every regression parameter.

Conclusion: all the coefficient of the regression parameters are different than 0.

Test 3:

Using the results of Test 2, reject Ho.

Conclusion: there is an interaction between the Show's Vote Count and the English Categorical variable.

Furthermore, the R-Square value now is 0.365412 which implies that removing 9 outliers from the data improved the linear relationship between the variables.

PROC FREQ DATA=STAT;

TABLE FILMLOC*RELSEASON;

RUN;

This proc produces a two-way frequency table of the categorical variables specified above.

Frequency Percent	Table of FILMLOC by RELSEASON				
	RELSEASON				
FILMLOC	Autumn	Spring	Summer	Winter	Total
1	892 34.16	545 20.87	504 19.30	549 21.03	2490 95.37
2	34 1.30	18 0.69	18 0.69	18 0.69	88 3.37
3	2 0.08	0 0.00	1 0.04	2 0.08	5 0.19
4 or more	2 0.08	3 0.11	0 0.00	2 0.08	7 0.27
Not given	6 0.23	7 0.27	2 0.08	6 0.23	21 0.80
Total	936 35.85	573 21.95	525 20.11	577 22.10	2611 100.00

DATA STAT2;

SET STAT;

LPOPULARITY = LOG(POPULARITY); * this line computes the natural log of the POPULARITY variable.

GROUPS = TRIM(FILMLOC) || ' - ' || RELSEASON; * this lines creates a new variable by combining the different values of the FILMLOC and RELSEASON.

RUN;

PROC UNIVARIATE DATA=STAT2;

TITLE 'Descriptive Statistics of LPOPULARITY';

VAR LPOPULARITY;

RUN;

Basic Statistical Measures			
Location		Variability	
Mean	3.440949	Std Deviation	0.91716
Median	3.314840	Variance	0.84119
Mode	2.881947	Range	8.95143
		Interquartile Range	1.10032

Quantiles (Definition 5)	
Level	Quantile
100% Max	8.80756
99%	6.30998
95%	5.13525
90%	4.65215
75% Q3	3.90984
50% Median	3.31484
25% Q1	2.80952
10%	2.44209
5%	2.22376
1%	1.84467
0% Min	-0.14387

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-0.143870	320	7.32119	824

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.168899	109	7.82125	1203
0.666803	1076	8.47376	340
0.811819	779	8.67691	956
1.383290	1138	8.80756	859

```
PROC GLM DATA=STAT2;
TITLE "Regression of Popularity with Show's Vote Count and the English Categorical Variable";
MODEL LPOPULARITY = SHOWVOTECOUNT | ENGLISH;
RUN;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	611.932516	203.977505	335.80	<.0001
Error	2607	1583.570020	0.607430		
Corrected Total	2610	2195.502535			

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.173810365	0.02863849	110.82	<.0001
SHOWVOTECOUNT	0.000405615	0.00002677	15.15	<.0001
ENGLISH	0.043140905	0.03566615	1.21	0.2266
SHOWVOTECOUN*ENGLISH	-0.000014701	0.00003027	-0.49	0.6273

```
PROC REG DATA=STAT2;
TITLE "Regression of Natural Log of Popularity with Show's Vote Count";
```

MODEL LPOPULARITY = SHOWVOTECOUNT;

OUTPUT OUT=RESIDUALS1 R=RES; * This line generates a new data set which contains the residuals (observed valued minus predicted values of the dependent variable) and predicted values obtained by performing simple regression.

RUN;

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	611.04176	611.04176	1006.15	<.0001
Error	2609	1584.46078	0.60731		
Corrected Total	2610	2195.50254			

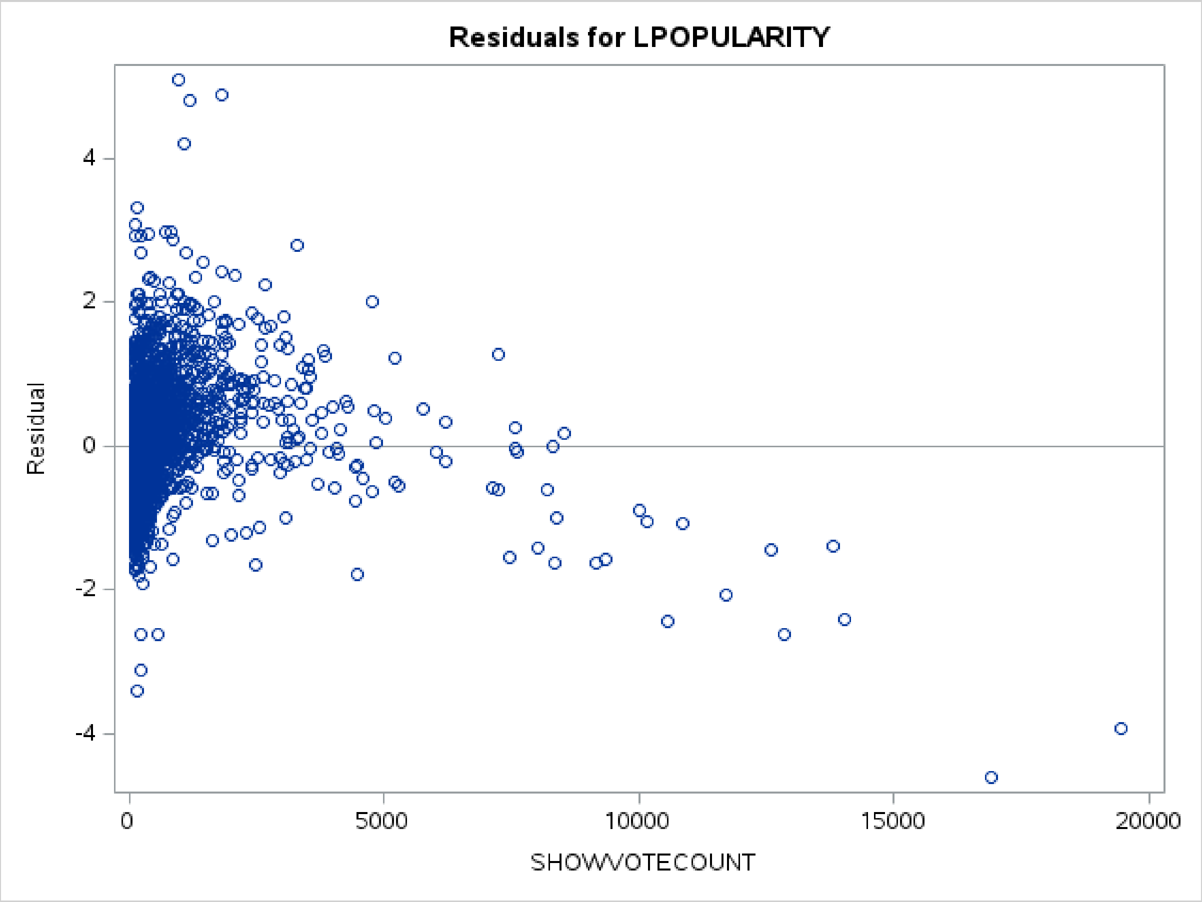
Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	3.20155	0.01702	188.15	<.0001
SHOWVOTECOUNT	1	0.00039516	0.00001246	31.72	<.0001

This linear regression model shows a better relationship than the one with the original data. The only that stills needs to be done is check that the data follows the 3 next assumptions of the normal linear regression model:

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

b. 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.



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

```
PROC UNIVARIATE DATA=RESIDUALS1 NORMAL;
TITLE 'Normal Test of RESIDUALS1';
VAR RES;
RUN;
```

Normal Test:

Ho: the residuals come from a normal distribution.

Ha: the residuals don't come from a normal distribution.

The same alpha value and decision rule of the previous tests will be used. Furthermore, since there are more than 2000 observations, the Kolmogorov-Smirnov test for normality is used.

Tests for Normality				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.051867	**Pr > D**	<0.0100

The p-value of this test is less than 0.01. Therefore, reject Ho.

Conclusion: 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.

ANOVA

The following procs will be used to perform one-way ANOVA Analysis using 3 different categorical variables.

Test for equality of means in the groups:

Ho: the mean values of each group are equal.

Ha: at least the mean of 1 group is different than the rest.

After doing the analysis, further tests will be done in each variable to see if the ANOVA model is a good fit for the data. The tests will confirm the next assumptions:

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

b. The errors of each group have a constant variance which will be 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.

Ho: the variances of the group are equal.

Ha: at least the variance of 1 group is different than the rest.

c. The errors come from a normal distribution which is checked that the residuals come from a normal distribution.

The same alpha value and decision rule of the previous tests will be used.

PROC GLM DATA=STAT2;

TITLE "ANOVA Analysis for SHOWVOTEAVERAGE Using FILMLOC Variable";

CLASS FILMLOC;

MODEL SHOWVOTEAVERAGE = FILMLOC;

LSMEANS FILMLOC / PDIFF ADJUST=TUKEY;

MEANS FILMLOC / TUKEY HOVTEST=BARTLETT;

OUTPUT OUT=RESIDUALS2 R=RES;

RUN;

Test for equality of means in the groups:

To solve the test, use the p-value in the first table of the output produce by this proc.

Class Level Information		
Class	**Levels**	**Values**
FILMLOC	5	1 2 3 4 or more Not given

Dependent Variable: SHOWVOTEAVERAGE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2.2640817	0.5660204	1.48	0.2049
Error	2606	995.1802844	0.3818804		
Corrected Total	2610	997.4443661			

The p-value is 0.2049 which is higher than alpha. Therefore, do not reject Ho.

Conclusion: the groups have a similar Show's Vote Average Mean.

Test for homogeneity of variance:

Bartlett's Test for Homogeneity of SHOWVOTEAVERAGE Variance			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
FILMLOC	4	7.0877	0.1313

The p-value produced by Bartlett's test is 0.1313 which is higher than alpha. Therefore, do not reject Ho.

Conclusion: the variances of the groups are equal.

```
PROC UNIVARIATE DATA=RESIDUALS2 NORMAL;
TITLE 'Normal Test of RESIDUALS2';
VAR RES;
RUN;
```

Test for Normality:

Tests for Normality				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.063329	**Pr > D**	<0.0100

The p-value is less than 0.01. Therefore, reject Ho.

Conclusion: the residuals don't come from a normal distribution.

The third assumption is violated. 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 this case.

```
PROC GLM DATA=STAT2;
TITLE "ANOVA Analysis of SHOWVOTEAVERAGE Using RELSEASON Variable";
CLASS RELSEASON;
MODEL SHOWVOTEAVERAGE = RELSEASON;
LSMEANS RELSEASON / PDIFF ADJUST=Tukey;
MEANS RELSEASON / TUKEY HOVTEST=BARTLETT;
OUTPUT OUT=RESIDUALS3 R=RES;
RUN;
```

Test for equality of means in the groups:

Class Level Information		
Class	**Levels**	**Values**
RELSEASON	4	Autumn Spring Summer Winter

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.2594595	0.0864865	0.23	0.8783
Error	2607	997.1849066	0.3825028		
Corrected Total	2610	997.4443661			

The p-value is 0.8783 which is higher than alpha. Therefore, do not reject Ho.

Conclusion: the groups have a similar Show's Vote Average Mean.

Test for homogeneity of variance:

Bartlett's Test for Homogeneity of SHOWVOTEAVERAGE Variance			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
RELSEASON	3	3.5479	0.3146

The p-value produced by Bartlett's test is 0.3146 which is higher than alpha. Therefore, do not reject Ho.

Conclusion: the variances of the groups are equal.

PROC UNIVARIATE DATA=RESIDUALS3 NORMAL;

TITLE 'Normal Test of RESIDUALS3';

VAR RES;

RUN;

Test for Normality:

Tests for Normality				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.05491	**Pr > D**	<0.0100

The p-value is less than 0.01. Therefore, reject Ho.

Conclusion: the residuals don't come from a normal distribution.

The third assumption is violated. 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.

PROC GLM DATA=STAT2;

TITLE "ANOVA Analysis of SHOWVOTEAVERAGE Using GROUPS Variable";

CLASS GROUPS;

MODEL SHOWVOTEAVERAGE = GROUPS;

LSMEANS GROUPS / PDIFF ADJUST=Tukey;

MEANS GROUPS / TUKEY HOVTEST=BARTLETT;

OUTPUT OUT=RESIDUALS4 R=RES;

RUN;

Test for equality of means in the groups:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	7.3982864	0.4351933	1.14	0.3082
Error	2593	990.0460797	0.3818149		
Corrected Total	2610	997.4443661			

The p-value is 0.3082 which is higher than alpha. Therefore, do not reject Ho.

Conclusion: the groups have a similar Show's Vote Average Mean.

Test for homogeneity of variance:

Bartlett's Test for Homogeneity of SHOWVOTEAVERAGE Variance			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
GROUPS	16	16.8358	0.3963

The p-value produced by Bartlett's test is 0.3963 which is higher than alpha. Therefore, do not reject Ho.

Conclusion: the variances of the groups are equal.

```
PROC UNIVARIATE DATA=RESIDUALS4 NORMAL;
TITLE 'Normal Test of RESIDUALS4';
VAR RES;
RUN;
```

Tests for Normality				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.057002	**Pr > D**	<0.0100

The p-value is less than 0.01. Therefore, reject Ho.

Conclusion: the residuals don't come from a normal distribution.

The third assumption is violated. 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.