Categorical Data Analysis

Table Of Contents

Categorical Data Analysis

<u>Dataset Description</u>:

A dataset with 218 records having 14 observations each is used to perform various tasks for the assignment. The dataset reflects on various parameters defining YouTube videos on different aspects. The dataset, in the CSV form as provided, is first uploaded on to the SAS platform and then imported into a custom created library named ‘You’ using the Proc Import statement. The imported dataset is named ‘youtube’. The above details are summarized below:

<u>S.no</u>	<u>Library Name</u>	<u>Dataset Name</u>	<u>no. of Rows</u>	<u>no. of columns</u>
1.	You	youtube	14	218

<u>Dataset Cleaning</u>:

The dataset is checked for missing values such as a character space ( ) or a period (.) using the ‘Describe missing data’ utility provided by the SAS platform. It has been identified that the dataset is not missing any values as such and a table has been provided below to support this.

Missing Data Frequencies

Legend: ., A, B, etc = Missing

VAR1	Frequency	Percent
Non-missing	218	100.00

VideoPublishTime	Frequency	Percent
Non-missing	218	100.00

ClicksEndPercent	Frequency	Percent
Non-missing	218	100.00

CommentsAdded	Frequency	Percent
Non-missing	218	100.00

Shares	Frequency	Percent
Non-missing	218	100.00

Dislikes	Frequency	Percent
Non-missing	218	100.00

Likes	Frequency	Percent
Non-missing	218	100.00

AveragePercentViewed	Frequency	Percent
Non-missing	218	100.00

AverageViewDuration	Frequency	Percent
Non-missing	218	100.00

Views	Frequency	Percent
Non-missing	218	100.00

WatchTimeHrs	Frequency	Percent
Non-missing	218	100.00

Subscribers	Frequency	Percent
Non-missing	218	100.00

Impressions	Frequency	Percent
Non-missing	218	100.00

ImpressionClick	Frequency	Percent
Non-missing	218	100.00

The dataset has also been checked for extreme values such as -999 to clean it further if identified any. The SAS ‘summary statistics’ utility has been used to analyse the dataset and there has been no trace of such values. The below table is provided to support the statement.

Variable	Mean	Std Dev	Minimum	Maximum	N
VAR1	115.1880734	66.6140022	1.0000000	236.0000000	218
VideoPublishTime	21892.96	148.0160613	21535.00	22146.00	218
ClicksEndPercent	7.5115596	3.0807766	0	15.7500000	218
CommentsAdded	1378.70	1894.48	0	14392.00	218
Shares	466.4082569	587.1371481	0	4116.00	218
Dislikes	1528.95	1676.90	0	9415.00	218
Likes	8209.25	7381.84	0	43733.00	218
AveragePercentViewed	30.5333486	6.1302160	8.8800000	57.0700000	218
AverageViewDuration	0.0670935	0.0210418	0.0080556	0.1388889	218
Views	1126416.48	1256118.22	2.0000000	8217897.00	218
WatchTimeHrs	83066.33	94304.08	0.1604000	565615.18	218
Subscribers	2093.01	2871.01	0	16518.00	218
Impressions	5963884.63	7521461.77	1438.00	46923937.00	218
ImpressionClick	11.7989450	3.3670977	0.1400000	20.2100000	218

A. <u>Binning Impressions</u>:

The Impressions column in the dataset has been binned using the PROC HPBIN statement further implementing the 'pseudo quantile' method. The pseudo quantile method divides the numeric data into quantiles with approximately equal values. The other parameters used are numbin, input and id. The numbin parameter decides on the number of quantiles to divide the numeric variable. The input parameter takes the numeric variable of interest and id is used to retain the other columns that are of interest in the output dataset. As per the requirement the numbin is provided with value 3 which divides the numeric variable into 3 quantiles with approximately equal values in each quantile. These categories have been marked as 1, 2, and 3 by default. The ranges are as specified below with quantile 1 having values less than 2267794.7017, 2nd quantile having values from the range 2267794.7017 to 5810443.3762 and 3rd quantile having values greater than or equal to 5810443.3762. Each quantile has approximately 73 values.

Mapping				
Variable	**Binned Variable**	**Range**	**Frequency**	**Proportion**
Impressions	**BIN_Impressions**	Impressions < 2267794.7017	73	0.33486239
		2267794.7017 <= Impressions < 5810443.3762	73	0.33486239
		5810443.3762 <= Impressions	72	0.33027523

The output dataset after performing binning for the impressions variable is named as 'bin' and the column impressions which has been binned is named as 'BIN_Impressions'.

Since the expected classification is to be labelled as low, medium, and high, a new custom format has been created using 'Proc Format' named as 'youcate'. This format labels all values with 1 as low, 2 as medium and 3 as high. The format on applying to the 'bin' dataset labels all 1st quantile elements as low, 2nd quantile as medium and 3rd quantile as high.

B. Binning other variables:

The dataset 'bin' has been analyzed to identify other continuous variables which can be binned using the 'Proc Hpbin' statement further implementing the 'pseudo quantile' method. All the other columns have been retained using the 'id' parameter. Numbin is assigned 3 which resulted in 3 quantiles each with labels 1, 2 and 3 representing 1st Quantile, 2nd Quantile and 3rd Quantile. The custom created format 'youcate' has been applied using the Proc Format to all the binned columns to label the 1 as low, 2 as medium and 3 as high. The following variables are identified on which the above tasks are performed:

i) Clicks.per.end.screen.element.shown.(%):

The variable has been renamed as 'ClicksEndPercent' for convenience. This variable is a continuous variable as it can absorb any value over the range of 0 to 100. The ranges are as specified below with quantile 1 having values less than 5.961375, 2nd quantile having values from the range 5.961375 to 8.941275 and 3rd quantile having values greater than or equal to 8.941275. Each quantile has approximately 73 values.

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
ClicksEndPercent	BIN_ClicksEndPercent	ClicksEndPercent < 5.961375	73	0.33486239
		5.961375 <= ClicksEndPercent < 8.941275	73	0.33486239
		8.941275 <= ClicksEndPercent	72	0.33027523

The resulted column is named as 'BIN_ClicksEndPercent'.

ii) Comments.Added :

The variable has been renamed as 'CommentsAdded' for convenience. This variable is regarded as a continuous variable as it can take any value. The variables is cut at a specific range with quantile 1 having values less than 561.288, 2nd quantile having values from the range 561.288 to 1098.1096 and 3rd quantile having values greater than or equal to 1098.1096. Each quantile has approximately 73 values.

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
CommentsAdded	BIN_CommentsAdded	CommentsAdded < 561.288	73	0.33486239
		561.288 <= CommentsAdded < 1098.1096	73	0.33486239
		1098.1096 <= CommentsAdded	72	0.33027523

The resulted column is named as 'BIN_CommentsAdded'.

iii) Shares:

This is a numeric continuous variable as there can be potentially any number of shares for a video. The variable is cut at a specific range with quantile 1 having values less than 167.1096, 2nd quantile having values from the range 167.1096 to 463.05 and 3rd quantile having values greater than or equal to 463.05. Each quantile has approximately 73 values.

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
Shares	BIN_Shares	Shares < 167.1096	73	0.33486239
		167.1096 <= Shares < 463.05	73	0.33486239
		463.05 <= Shares	72	0.33027523

The resulted column is named as 'BIN_Shares'.

iv) Dislikes:

This is a continuous variable as it can assume any value for the count of dislikes. The variable is cut at a specific range with quantile 1 having values less than 593.145, 2nd quantile having values from the range 593.145 to 1531.8205 and 3rd quantile having values greater than or equal to 1531.8205. Each quantile has approximately 73 values.

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
Dislikes	BIN_Dislikes	Dislikes < 593.145	73	0.33486239
		593.145 <= Dislikes < 1531.8205	73	0.33486239
		1531.8205 <= Dislikes	72	0.33027523

The resulted column is named as 'BIN_Dislikes'.

v) Likes:

Just like dislikes, this is also a continuous variable which can be binned into discrete values. The variable is divided into ranges with quantile 1 having values less than 4333.9403, 2nd quantile having values from the range 4333.9403 to 9035.2378 and 3rd quantile having values greater than or equal to 9035.2378. Each quantile has approximately 73 values.

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
Likes	BIN_Likes	Likes < 4333.9403	73	0.33486239
		4333.9403 <= Likes < 9035.2378	73	0.33486239
		9035.2378 <= Likes	72	0.33027523

The resulted column is named as ‘BIN_Likes’.

vi) Average.percentage.viewed.(%):

The variable has been renamed as ‘AveragePercentViewed’ for convenience. This variable is a continuous variable as it can absorb any value over the range of 0 to 100. The ranges are as specified below with quantile 1 having values less than 29.153533, 2nd quantile having values from the range 29.153533 to 32.724412 and 3rd quantile having values greater than or equal to32.724412. Each quantile has approximately 73 values.

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
AveragePercentViewed	BIN_AveragePercentViewed	AveragePercentViewed < 29.153533	73	0.33486239
		29.153533 <= AveragePercentViewed < 32.724412	73	0.33486239
		32.724412 <= AveragePercentViewed	72	0.33027523

The resulted column is named as ‘BIN_AveragePercentViewed’.

vii) Views:

This is a continuous variable as there can be any number of views for a video. The variable is cut at a specific range with quantile 1 having values less than 425688.961, 2nd quantile having values from the range 425688.961 to 1206388.986 and 3rd quantile having values greater than or equal to 1206388.986. Each quantile has approximately 73 values.

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
Views	BIN_Views	Views < 425688.961	73	0.33486239
		425688.961 <= Views < 1206388.986	73	0.33486239
		1206388.986 <= Views	72	0.33027523

The resulted column is named as ‘BIN_Views’.

viii) Watch.time.(hours):

The variable has been renamed as ‘WatchTimeHrs’ for convenience. This variable is a continuous variable with time measured in hours and hench is chosen to convert into discrete values. The ranges are as specified below with quantile 1 having values less than 26810.312547, 2nd quantile having values from the range 26810.312547 to 92761.024369 and 3rd quantile having values greater than or equal to 92761.024369. Each quantile has approximately 73 values.

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
WatchTimeHrs	BIN_WatchTimeHrs	WatchTimeHrs < 26810.312547	73	0.33486239
		26810.312547 <= WatchTimeHrs < 92761.024369	73	0.33486239
		92761.024369 <= WatchTimeHrs	72	0.33027523

The resulted column is named as 'BIN_WatchTimeHrs'.

ix) <u>Subscribers</u>:

This is a continuous variable as there can be any number of subscribers. The variable is cut at a specific range with quantile 1 having values less than 518.6652, 2nd quantile having values from the range 518.6652 to 1815.3282 and 3rd quantile having values greater than or equal to 1815.3282. Each quantile has approximately 73 values.

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
Subscribers	BIN_Subscribers	Subscribers < 518.6652	73	0.33486239
		518.6652 <= Subscribers < 1815.3282	73	0.33486239
		1815.3282 <= Subscribers	72	0.33027523

The resulted column is named as 'BIN_Subscribers'.

x) <u>Impressions.click-through.rate.(%)</u>:

The variable has been renamed as 'ImpressionClick' for convenience. This variable is a continuous variable as it can absorb any value over the range of 0 to 100. The ranges are as specified below with quantile 1 having values less than 11.150402, 2nd quantile having values from the range 11.150402 to 13.380179 and 3rd quantile having values greater than or equal to 13.380179. Each quantile has approximately 73 values.

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
ImpressionClick	BIN_ImpressionClick	ImpressionClick < 11.150402	73	0.33486239
		11.150402 <= ImpressionClick < 13.380179	73	0.33486239
		13.380179 <= ImpressionClick	72	0.33027523

The resulted column is named as 'BIN_ImpressionClick'.

C. <u>Categorical Relationships</u>:

The Categorical relationship between two variables is assessed using the Chi square test. This test helps in understanding the association between two categorical variables. The test is performed using 'Proc Freq' statement further implementing 'chisq' method. Upon interpreting the results, we can assert on the relationship of the variables. A few meaningful pairs have been selected and analyzed from the dataset whose results are presented below:

i) <u>BIN_Impressions **vs** BIN_Views</u>:

BIN_Impressions variable defines the number of times a thumbnails for each video was shown to YouTube viewers, whereas BIN_Views gives the number of viewers that watched the video.

Upon running the chi square test and generating the two-way table the following results are presented.

Frequency
Expected
Percent
Row Pct
Col Pct

Table of BIN_Impressions by BIN_Views

	BIN_Views			
BIN_Impressions	1	2	3	Total
1	68 24.445 31.19 93.15 93.15	5 24.445 2.29 6.85 6.85	0 24.11 0.00 0.00 0.00	73 33.49
2	5 24.445 2.29 6.85 6.85	64 24.445 29.36 87.67 87.67	4 24.11 1.83 5.48 5.56	73 33.49
3	0 24.11 0.00 0.00 0.00	4 24.11 1.83 5.56 5.48	68 23.78 31.19 94.44 94.44	72 33.03
Total	73 33.49	73 33.49	72 33.03	218 100.00

Statistics for Table of BIN_Impressions by BIN_Views

Statistic	DF	Value	Prob
Chi-Square	4	336.5431	<.0001
Likelihood Ratio Chi-Square	4	344.7443	<.0001
Mantel-Haenszel Chi-Square	1	190.8973	<.0001
Phi Coefficient		1.2425	
Contingency Coefficient		0.7790	
Cramer's V		0.8786	

Sample Size = 218

Upon analyzing the two-way table, it can be observed that the expected count is greater than 5 for every cell and hence we can reject the possibility of overlapping. From the chi square row, it can be identified that the statistic value 336.5431 is strictly higher than the DP of 9.49 (DF = 4) and p value is smaller than the assumed alpha value (0.05). So having enough evidence to reject null hypothesis, we can assert that the two variables are dependent on each other.

So, there is an association between BIN_Impressions and BIN_Views which means more the Impressions shown for each video more the chances of an increase in the view count.

ii) <u>BIN_Shares **vs** BIN_Likes</u>:

BIN_Shares variable defines the shares for each video, whereas BIN_Likes gives the number of likes for each video.

Upon running the chi square test and generating the two-way table the following results are presented.

The FREQ Procedure

Frequency Expected Percent Row Pct Col Pct

Table of BIN_Shares by BIN_Likes				
	BIN_Likes			
BIN_Shares	1	2	3	Total
1	63 24.445 28.90 86.30 86.30	9 24.445 4.13 12.33 12.33	1 24.11 0.46 1.37 1.39	73 33.49
2	10 24.445 4.59 13.70 13.70	54 24.445 24.77 73.97 73.97	9 24.11 4.13 12.33 12.50	73 33.49
3	0 24.11 0.00 0.00 0.00	10 24.11 4.59 13.89 13.70	62 23.78 28.44 86.11 86.11	72 33.03
Total	73 33.49	73 33.49	72 33.03	218 100.00

Statistics for Table of BIN_Shares by BIN_Likes

Statistic	DF	Value	Prob
Chi-Square	4	240.2560	<.0001
Likelihood Ratio Chi-Square	4	246.1451	<.0001
Mantel-Haenszel Chi-Square	1	158.6947	<.0001
Phi Coefficient		1.0498	
Contingency Coefficient		0.7241	
Cramer's V		0.7423	

Sample Size = 218

On analyzing the two-way table, it can be observed that the expected count is greater than 5 for every cell and hence we can reject the possibility of overlapping. From the chi square row, it can be identified that the statistic value 240.2560 is strictly higher than the DP of 9.49 (DF = 4) and p value is smaller than the assumed alpha value (0.05). So having enough evidence to reject null hypothesis, we can assert that the two variables are dependent on each other.

So, there is a significant association between BIN_Shares and BIN_Likes and it is clear from the two-way analysis table that higher the shares for a video, more the likes.

iii) <u>BIN_Shares **vs** BIN_Views</u>:

BIN_Shares variable defines the shares for each video, whereas BIN_Views gives the number of views for each video.

Upon running the chi square test and generating the two-way table the following results are presented.

The FREQ Procedure

Frequency Expected Percent Row Pct Col Pct

Table of BIN_Shares by BIN_Views

	BIN_Views			
BIN_Shares	1	2	3	Total
1	65 24.445 29.82 89.04 89.04	8 24.445 3.67 10.96 10.96	0 24.11 0.00 0.00 0.00	73 33.49
2	8 24.445 3.67 10.96 10.96	59 24.445 27.06 80.82 80.82	6 24.11 2.75 8.22 8.33	73 33.49
3	0 24.11 0.00 0.00 0.00	6 24.11 2.75 8.33 8.22	66 23.78 30.28 91.67 91.67	72 33.03
Total	73 33.49	73 33.49	72 33.03	218 100.00

Statistics for Table of BIN_Shares by BIN_Views

Statistic	DF	Value	Prob
Chi-Square	4	288.6420	<.0001
Likelihood Ratio Chi-Square	4	296.7302	<.0001
Mantel-Haenszel Chi-Square	1	177.1183	<.0001
Phi Coefficient		1.1507	
Contingency Coefficient		0.7548	
Cramer's V		0.8136	

Sample Size = 218

On analyzing the two-way table, it can be observed that the expected count is greater than 5 for every cell and hence we can reject the possibility of overlapping. From the chi square row, it can be identified that the statistic value of 288.6420 is significantly higher than the DP of 9.49 (DF = 4) and p value is smaller than the assumed alpha value (0.05). So having enough evidence to reject null hypothesis, we can assert that the two variables are dependent on each other.

So, there is a significantly strong association between the two variables, and it is clear from the two-way analysis table that higher the shares for a video, more the views the video has.

iv) <u>BIN_WatchTimeHrs **vs** BIN_Subscribers</u>:

BIN_WatchTimeHrs variable defines the total number of hours watched for each video, whereas BIN_Subscribers gives the number of subscribers for each video. Upon running the chi square test and generating the two-way table the following results are presented.

The FREQ Procedure

Frequency Expected Percent Row Pct Col Pct

Table of BIN_WatchTimeHrs by BIN_Subscribers

	BIN_Subscribers			
BIN_WatchTimeHrs	1	2	3	Total
1	61 24.445 27.98 83.56 83.56	12 24.445 5.50 16.44 16.44	0 24.11 0.00 0.00 0.00	73 33.49
2	12 24.445 5.50 16.44 16.44	51 24.445 23.39 69.86 69.86	10 24.11 4.59 13.70 13.89	73 33.49
3	0 24.11 0.00 0.00 0.00	10 24.11 4.59 13.89 13.70	62 23.78 28.44 86.11 86.11	72 33.03
Total	73 33.49	73 33.49	72 33.03	218 100.00

Statistics for Table of BIN_WatchTimeHrs by BIN_Subscribers

Statistic	DF	Value	Prob
Chi-Square	4	222.3484	<.0001
Likelihood Ratio Chi-Square	4	236.0481	<.0001
Mantel-Haenszel Chi-Square	1	156.1453	<.0001
Phi Coefficient		1.0099	
Contingency Coefficient		0.7106	
Cramer's V		0.7141	

Sample Size = 218

On analyzing the two-way table, it can be observed that the expected count is greater than 5 for every cell and hence we can reject the possibility of overlapping. From the chi square row, it can be identified that the statistic value of 222.3484 is significantly higher than the DP of 9.49 (DF = 4) and p value is smaller than the assumed alpha value (0.05). So having enough evidence to reject null hypothesis, we can assert that the two variables are dependent on each other.

We can conclude that there is a significant association between the two variables, and it is clear from the two-way analysis table that higher the subscribers for a video, more the watch time hours.

v) <u>BIN_ClicksEndPercent **vs** BIN_Views</u>:

BIN_ClicksEndPercent variable depicts the percentage of viewers who selected the end screen, whereas BIN_Views gives the total number of viewers for each video. Upon running the chi square test and generating the two-way table the following results are presented.

The FREQ Procedure

Frequency
Expected
Percent
Row Pct
Col Pct

Table of BIN_ClicksEndPercent by BIN_Views

BIN_ClicksEndPercent	BIN_Views 1	2	3	Total
1	26 24.445 11.93 35.62 35.62	30 24.445 13.76 41.10 41.10	17 24.11 7.80 23.29 23.61	73 33.49
2	22 24.445 10.09 30.14 30.14	27 24.445 12.39 36.99 36.99	24 24.11 11.01 32.88 33.33	73 33.49
3	25 24.11 11.47 34.72 34.25	16 24.11 7.34 22.22 21.92	31 23.78 14.22 43.06 43.06	72 33.03
Total	73 33.49	73 33.49	72 33.03	218 100.00

Statistics for Table of BIN_ClicksEndPercent by BIN_Views

Statistic	DF	Value	Prob
Chi-Square	4	8.9233	0.0630
Likelihood Ratio Chi-Square	4	9.2553	0.0550
Mantel-Haenszel Chi-Square	1	2.3210	0.1276
Phi Coefficient		0.2023	
Contingency Coefficient		0.1983	
Cramer's V		0.1431	

Sample Size = 218

On analyzing the two-way table, it can be observed that the expected count is greater than 5 for every cell and hence we can reject the possibility of overlapping. From the chi square row, it can be identified that the statistic value of 8.9233 is lower than the DP of 9.49 (DF = 4). So, there is no enough evidence to reject null hypothesis, hence we can conclude that they are significantly independent.

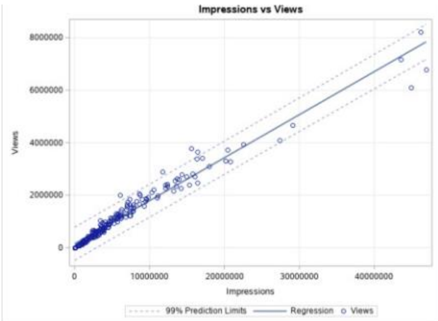
So, we can conclude that despite the increase in the total viewers, the share of viewers who selected end screen elements is not changing and is arbitrary in relation to the viewer count.

D. <u>Categorical Analysis VS Correlation</u>:

The Correlation analysis is first conducted using the scatter plot after which we perform pearson correlation test to assert the relation between the two numeric variables. Scatter plot is implemented using the ‘Proc SGPLOT’ method and Pearson correlation is done by implementing ‘Proc CORR’.

i) <u>Impressions **vs** Views</u>:

<u>Scatter plot</u>:



The above Scatter plot shows a highly strong positive linear relation ,nearly perfect, between the two variables which mean more the impressions more the views.

Impressions vs Views

1 With Variables:	Views
1 Variables:	Impressions

Pearson Correlation Coefficients, N = 218 Prob > |r| under H0: Rho=0	
	Impressions
Views	0.98155 <.0001

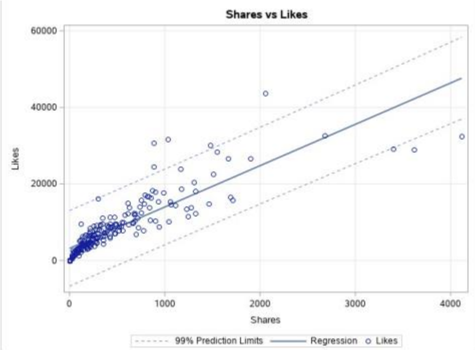
The correlation coefficient (r) being 0.98155 which is positive and almost close to 1 in magnitude tells us that the relation between the two variables is highly positive.

<u>Comment</u>:

The chi square and the correlation analysis point to the same results about the association between both the variables which is they are strongly associated.

ii) <u>Shares **vs** Likes</u>:

Scatter plot:



The above Scatter plot shows a high positive linear relation between the two variables which mean more the shares to a video more the chances of getting likes.

Shares vs Likes

1 With Variables:	Likes
1 Variables:	Shares

Pearson Correlation Coefficients, N = 218 Prob > |r| under H0: Rho=0	
	Shares
Likes	0.85896 <.0001

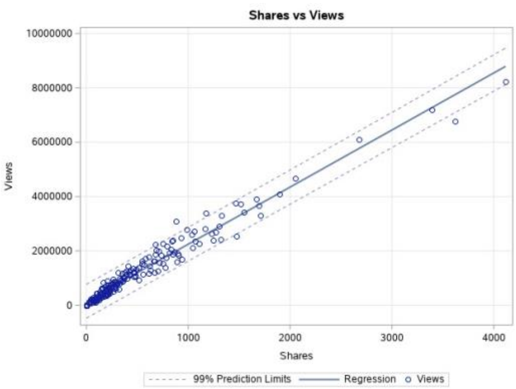
The correlation coefficient (r) being 0.85896 which is positive and almost close to 0.9 in magnitude tells us that the relation between the two variables is highly positive.

Comment:

The chi square and the correlation analysis point to the same results about the association between both the variables which is they are highly associated.

iii) Shares **vs** Views:

Scatter plot:



The above Scatter plot shows a highly strong positive linear relation, nearly perfect, between the two variables which mean more the shares for a video, more the views.

Shares vs Views

1 With Variables:	Views
1 Variables:	Shares

Pearson Correlation Coefficients, N = 218 Prob > |r| under H0: Rho=0	
	Shares
Views	0.98193 <.0001

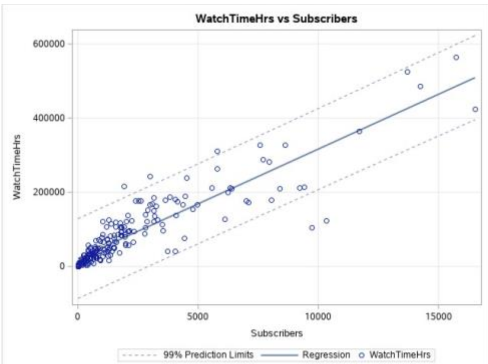
The correlation coefficient (r) being 0.98193 which is positive and almost close to 1 in magnitude tells us that the relation between the two variables is highly positive.

Comment:

The chi square and the correlation analysis point to the same results about the association between both the variables which is they are strongly dependent.

iv) WatchTimeHrs **vs** Subscribers:

Scatter plot:



The above Scatter plot, although a little distributed, shows a strong positive linear relation between the two variables which mean more the subscribers to a video more the WatchTimeHrs.

WatchTimeHrs vs Subscribers

1 With Variables:	WatchTimeHrs
1 Variables:	Subscribers

Pearson Correlation Coefficients, N = 218 Prob > |r| under H0: Rho=0	
	Subscribers
WatchTimeHrs	0.89948 <.0001

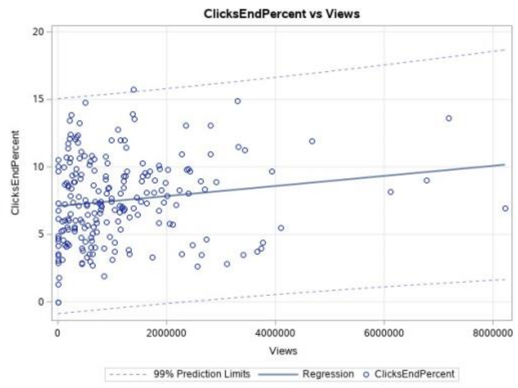
The correlation coefficient (r) being 0.89948 which is positive and almost close to 0.9 in magnitude tells us that the relation between the two variables is highly positive.

Comment:

The chi square and the correlation analysis point to the same results about the association between both the variables which is they are dependent on each other.

v) ClicksEndPercent **vs** Views:

Scatter plot:



The above Scatter plot shows no correlation as the graph is widely distributed with no pattern between the two variables.

ClicksEndPercent vs Views

1 With Variables:	Views
1 Variables:	ClicksEndPercent

Pearson Correlation Coefficients, N = 218 Prob > |r| under H0: Rho=0	
	ClicksEndPercent
Views	0.15243 0.0244

The correlation coefficient (r) being 0.15243 which is almost close to 0 indicates that there is no correlation and confirms our interpretation from the scatter plot.

Comment:

The chi square and the correlation analysis point to the same results about the association between both the variables which is that they are not dependent on each other and act as individual variables when comes to relation between them.

E. Variable Selection:

- From the above analysis of chi square and correlation, it is to some extent clear that Impressions is playing a key role. More the impressions for a video, more the views for it. Since views can directly impact the likes, shares and comments for any video, we can safely assume that Impressions is a predictor variable, and views, likes, shares and comments are dependent variables.
- Subscribers can be used as a predictor to predict the watch time hours. Since subscribers are the people who are interested in the content posted by the channel and receive notifications of every upload, it is very likely that subscribers watch more percent of the videos than non-subscribers. So having high number of subscribers positively impacts the watch time hours of the video and increase the views as they are notified of uploads.
- Shares to a video can be key in getting more views from non-subscribers. The content once shared can bring in new views and open doors to more shares. So more shares can increase the views thus leading to more likes/dislikes and comments.

Below table is the summary of the above information:

Predictors	Dependents
1. Impressions	Views
2. Subscribers	views, Watch Time Hours
3. Shares	Views

F. Linear Regression Model:

The below linear Regression Model is build using multiple predictor variables. The dependent variable predicted is views and the predictors used are Impressions, Subscribers and Shares.

The following model is built on SAS by implementing ‘proc reg’. The following results are generated on implementing the logic.

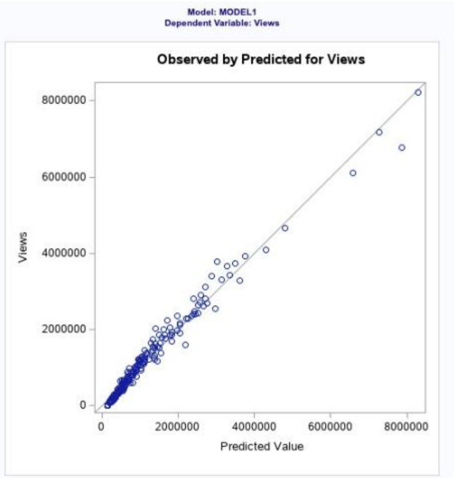
Model: MODEL1
Dependent Variable: Views

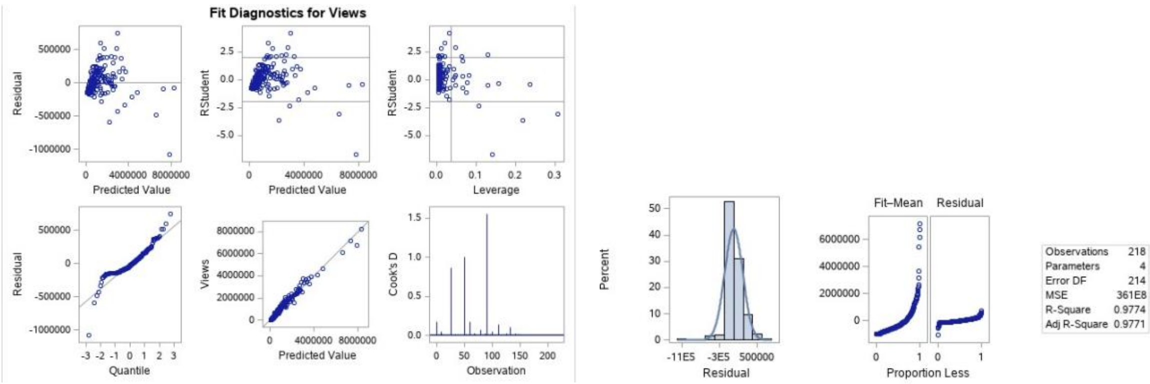
Number of Observations Read	218
Number of Observations Used	218

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3.34661E14	1.115537E14	3088.77	<.0001
Error	214	7.728797E12	36115872749		
Corrected Total	217	3.423898E14			

Root MSE	190042	R-Square	0.9774
Dependent Mean	1126416	Adj R-Sq	0.9771
Coeff Var	16.87136		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	138605	16506	8.40	<.0001
Impressions	1	0.06434	0.00896	7.18	<.0001
Subscribers	1	52.87601	13.51016	3.91	0.0001
Shares	1	1057.90805	99.85069	10.59	<.0001





From the above results, it is identified that the p value (0.0001) is less than the alpha value of 0.05 which indicates that there is no overlapping the model is explaining a lot of variability. Also, the R-square value which stands at 0.9774, almost close to 1, tells us that our model is predicting 97.74% change in the dependent variable when predictors are varied.

On looking at the graph generated, it is understood that there is a positive linear relation between the predictors and dependent variables which declares our model to be very accurate.

The assumptions of data being normal and independent from each other can be confirmed from the residual graphs hence making the linear model valid.

G. Recommendations:

The Recommendations below are provided in favor of boosting the views for any YouTube video irrespective of the content offered. A thorough analysis of the YouTube dataset and the regression model led to the following recommendations:

- Increase the rate of impressions:
  More number of times impressions are shown for a video, higher the chances of a viewer clicking on the thumbnail thus increasing the view count.
- Attractive thumbnail for impressions:
  Keeping attractive thumbnails can make the viewers click on it thus increasing the views.
- Active content uploading and Encourage subscriptions:
  Uploading content frequently helps random viewers to become subscribers. Views from subscribers are not dependent on the random impressions as they are notified of all the content thus eliminating arbitrary factor. Also encourage viewers to subscribe by requesting them in videos.
- Promote content on multiple platforms for more shares:
  Using more social platforms to promote the content can bring in more reach thus leading to more views.