PURPOSE

This report provides a detailed explanation of Categorical Data Analysis of YouTube data collected from a Youtuber's Account. This report is intended for:

- Exploring the Datasets.
- Understanding the data.
- Screening/Cleaning the data
- Binning the Numerical Data to Create Categorical Variables
- Determining relationships between categorical variables
- Comparing Relationships of Numerical and Corresponding Categorical Variables
- Selection of Predictors and Response Variable for Prediction Model
- Linear Regression Modelling
- Interpretations and Recommendations to Administrator

Data Description

- The dataset contains records of metrics related to videos uploaded on a Youtuber's Account.
- Most common and influential metrics are present in dataset such as number of Likes, Shares, Comments etc. on the video along with the Publish date.
- The data has numerical values for each metric.

Overview

The analysis starts with exploring the datasets. The number and the type of variables along with the number of observations are explored to get an overview of the dataset. Since the dataset is clean, no steps are preformed for pre-processing the data. Next, all numerical variables of the dataset are converted into Categorical variables by defining 3 ordered levels 'Low', 'Medium' and 'High' and categorizing data using Fixed-Frequency Binning.

After obtaining the Categorical variables, Two-Way Tables are constructed, and Chi-Square test is performed to determine the existence of relationships between different variables. Stacked bar graphs are plotted to show equivalence of distributions and describe independence.

A comparison is then performed between correlation obtained from Numerical data with the results of Chi-Square tests to determine any change in relationships based on tests. Lastly, Linear regression is performed by first selecting predictors and a response variable. Based on results, recommendations are provided to the administrator to improve performance of account.

Analyzing Dataset:

Table Attributes and Data:

Data Structure for Youtube Videos Data

Data Set Name	A3.YOUTUBE	Observations	218
Member Type	DATA	Variables	13
Engine	V9	Indexes	0
Created	27/02/2021 19:51:27	Observation Length	104
Last Modified	27/02/2021 19:51:27	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Variables in Creation Order

#	Variable	Type	Len	Format	Informat
1	Video_publish_time	Num	8	DDMMYY10.	DDMMYY10.
2	Clicks_per_end_screen_element_sh	Num	8	BEST12.	BEST32.
3	Comments_added	Num	8	BEST12.	BEST32.
4	Shares	Num	8	BEST12.	BEST32.
5	Dislikes	Num	8	BEST12.	BEST32.
6	Likes	Num	8	BEST12.	BEST32.
7	Average_percentage_viewed____	Num	8	BEST12.	BEST32.
8	Average_view_duration	Num	8	BEST12.	BEST32.
9	Views	Num	8	BEST12.	BEST32.
10	Watch_time__hours_	Num	8	BEST12.	BEST32.
11	Subscribers	Num	8	BEST12.	BEST32.
12	Impressions	Num	8	BEST12.	BEST32.
13	Impressions_click_through_rate__	Num	8	BEST12.	BEST32.

YouTube Data First 10 Observations

S.No.	Video_publish_time	Clicks_per_end_screen_element_sh	Comments_added	Shares	Dislikes	Likes	Average_percentage_viewed____
1	25/04/2020	2.68	14392	1476	3502	30097	21.34
2	26/04/2020	11.27	9640	1544	4731	28341	29.33
3	06/04/2020	7.12	8524	196	505	5353	28.03
4	22/04/2020	4.75	8357	321	884	9474	32.1
5	15/04/2020	8.14	8190	781	2185	14919	33.99
6	03/04/2020	8.94	8095	1034	3233	31784	41.64
7	23/04/2020	11.94	7260	2053	5731	43733	33.59
8	09/04/2020	13.08	7222	1161	3577	23979	31.25
9	27/03/2020	8.87	5857	1305	3435	20409	25.5
10	30/03/2020	7.07	5064	611	1630	15064	31.12

Average_view_duration	Views	Watch_time__hours_	Subscribers	Impressions	Impressions_click_through_rate__
0.110833333	2556273	283662.3563	7938	13405459	10.92
0.062777778	3425226	215723.1428	9410	17022958	14.36
0.051666667	375546	19504.5802	885	2523477	8.48
0.0625	603314	37711.324	1705	3862723	9.12
0.071111111	1758474	125359.3992	3307	8278024	14.42
0.047222222	2618929	124216.5309	10333	13918672	12.86
0.078055556	4663177	364179.2895	11691	29094097	11.19
0.075833333	2806856	212984.1362	6334	15758684	12.42
0.043888889	2898397	127720.0356	6088	11742075	16.99
0.065	1179001	76846.7102	4407	6792147	11.68

From the above tables about data structure, it can be observed that:

- Number of Variables: 13
- Number of Observations: 218
- Numerical Variables: 13 {All}
- Continuous Numerical Variables:
    - Clicks_per_end_screen_element_shown(%)
    - Average_percentage_viewed_(%)
    - Average_View_Duration
    - Watch_time_hours
    - Impressions_click_through_rate(%)
- Discreate Numerical Variable:
    - Comments_Added
    - Shares
    - Dislikes
    - Likes
    - Views
    - Subscribers
    - Impressions

This provides an overview of data and its structure. To dive deeper into the data and issues related, further analysis is done for numerical variables.

Analyzing Numerical Variables

Descriptive Statistics for Numeric Variables

Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev	Skewness	Kurtosis
Clicks_per_end_screen_element_sh	218	0	0	7.5115596	7.3450000	15.7500000	3.0807766	0.1749968	-0.4250303
Comments_added	218	0	0	1378.70	792.0000000	14392.00	1894.48	3.2896969	14.0689167
Shares	218	0	0	466.4082569	266.0000000	4116.00	587.1371481	3.0656921	12.9744966
Dislikes	218	0	0	1528.95	964.0000000	9415.00	1676.90	2.1780995	5.5461545
Likes	218	0	0	8209.25	6400.00	43733.00	7381.84	1.7710307	3.8527839
Average_percentage_viewed____	218	0	8.8800000	30.5333486	31.0000000	57.0700000	6.1302160	-0.2288100	3.0348628
Average_view_duration	218	0	0.0080556	0.0670935	0.0683333	0.1388889	0.0210418	-0.0541054	0.7732265
Views	218	0	2.0000000	1126416.48	723098.50	8217897.00	1256118.22	2.5031632	8.7745029
Watch_time__hours_	218	0	0.1604000	83066.33	49911.58	565615.18	94304.08	2.2214771	6.5875655
Subscribers	218	0	0	2093.01	1090.50	16518.00	2871.01	2.5588884	7.5185246
Impressions	218	0	1438.00	5963884.63	3706482.50	46923937.00	7521461.77	3.1337611	12.6773045
Impressions_click_through_rate__	218	0	0.1400000	11.7989450	12.0500000	20.2100000	3.3670977	-1.0354543	1.9493008

Minimum and Maximum Dates

Date variable	Minimum date	Maximum date
Video_publish_time	17/12/2018	19/08/2020

The descriptive statistics of the Numerical Variables has been calculated using the 'proc means' command in SAS. The above table provides the following information:

- There is no missing data in the dataset. Also, there is no invalid data like -999 etc. which needs to be addressed.
- Majority of the variables are positively skewed (67%). Three variables Clicks_per_end_Screen_element_Sh, Average_Percentage_Viewed__ and Average_View_Duration are approximately normally distributed. On the other hand, the Impressions_click_through_rate_ is negatively skewed.
- Nine of the variables are oridinal with discreate values describing counts while rest three i.e., Clicks_per_end_Screen_element_Sh, Average_Percentage_Viewed__ and Impressions_click_through_Rate__ are percentages with continuous values.

Binning:

Discretization or Binning is used for transforming numerical variables into categorical features. These features can be thought of as bins into which the raw numeric values are binned or grouped into. Each bin represents a specific degree of intensity and hence a specific range of continuous numeric values fall into it. The problem of working with numeric features is that the distribution of values in these features might be skewed along with varying range of values in these features. For instance, view counts of specific music videos could be abnormally large, and some could be really small. Directly using these features can cause a lot of issues and adversely affect the model.

There are two types of binning:

- Fixed-Width Binning: The width of bins is fixed irrespective of number of values in the bin.
- Adaptive Binning: Bin width depends on the data and its variations.

Quantile or Rank based binning is a good strategy which is used for adaptive binning in this analysis. Quantiles are specific values or cut-points which help in partitioning the continuous valued distribution of a specific numeric field into discrete contiguous bins or intervals. Thus, q-Quantiles help in partitioning a numeric attribute into q equal partitions.

Here, 3-Quantile binning is being used to divide range of values for different values into 3 categories by applying 2 cut points. These categories are:

- Low: Consisting of lower values below first tertile
- Medium: Consisting of moderate values between first and second tertile.
- High: Consisting of higher values beyond second tertile

The binning is achieved by using PROC RANK procedure in SAS. A new dataset is thus, created from original dataset by binning 12 of the 13 variables (leaving Publish Time since its date). Since, Rank procedure gives numeric values opposed to what is required in this report, a custom Format is used using Proc Format to map these numeric values of 0,1 and 2 to Low, Medium and High respectively.

Binned Youtube Data First 10 Observations

S.No.	Video_publish_time	Average Percentage Viewed	Average View Duration	Clicks/End Screen Element Shown	Comments Added	Dislikes	Impressions	Impressions CTR	Likes	Shares	Subscribers	Views	Watch Time Hours
1	25/04/2020	Low	High	Low	High	High	High	Low	High	High	High	High	High
2	26/04/2020	Med	Med	High	High	High	High	High	High	High	High	High	High
3	06/04/2020	Low	Low	Med	High	Low	Med	Low	Med	Med	Med	Low	Low
4	22/04/2020	Med	Med	Low	High	Med	Med	Low	High	Med	Med	Med	Med
5	15/04/2020	High	Med	Med	High	High	High	High	High	High	High	High	High
6	03/04/2020	High	Low	High	High	High	High	Med	High	High	High	High	High
7	23/04/2020	High	High	High	High	High	High	Med	High	High	High	High	High
8	09/04/2020	Med	High	High	High	High	High	Med	High	High	High	High	High
9	27/03/2020	Low	Low	Med	High	High	High	High	High	High	High	High	High
10	30/03/2020	Med	Med	Med	High	High	High	Med	High	High	High	Med	Med

N = 10

Frequency of Binned Variables

Rank for Variable Average_percentage_viewed____

Average_percentage_viewed____	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	72	33.03	72	33.03
Med	73	33.49	145	66.51
High	73	33.49	218	100.00

Rank for Variable Average_view_duration

Average_view_duration	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	72	33.03	72	33.03
Med	74	33.94	146	66.97
High	72	33.03	218	100.00

Rank for Variable Clicks_per_end_screen_element_sh

Clicks_per_end_screen_element_sh	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	72	33.03	72	33.03
Med	73	33.49	145	66.51
High	73	33.49	218	100.00

Rank for Variable Comments_added

Comments_added	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	72	33.03	72	33.03
Med	73	33.49	145	66.51
High	73	33.49	218	100.00

Rank for Variable Dislikes

Dislikes	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	72	33.03	72	33.03
Med	73	33.49	145	66.51
High	73	33.49	218	100.00

Rank for Variable Impressions

Impressions	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	72	33.03	72	33.03
Med	73	33.49	145	66.51
High	73	33.49	218	100.00

Rank for Variable Impressions_click_through_rate__

Impressions_click_through_rate__	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	73	33.49	73	33.49
Med	72	33.03	145	66.51
High	73	33.49	218	100.00

Rank for Variable Likes

Likes	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	72	33.03	72	33.03
Med	73	33.49	145	66.51
High	73	33.49	218	100.00

Rank for Variable Shares				
Shares	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	72	33.03	72	33.03
Med	73	33.49	145	66.51
High	73	33.49	218	100.00

Rank for Variable Subscribers				
Subscribers	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	72	33.03	72	33.03
Med	73	33.49	145	66.51
High	73	33.49	218	100.00

Rank for Variable Views				
Views	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	72	33.03	72	33.03
Med	73	33.49	145	66.51
High	73	33.49	218	100.00

Rank for Variable Watch_time__hours_				
Watch_time__hours_	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	72	33.03	72	33.03
Med	73	33.49	145	66.51
High	73	33.49	218	100.00

The above tables describe the structure of dataset after binning. The following can be observed:

- The dataset now consists of 11 Categorical variables with categories as ‘LOW’, ‘MED’ and ‘HIGH’. The data is a time series data which each observation corresponding to a date called Publish Date.
- Since, the binning is performed using Adaptive/Quantile Binning, each category is equally likely in the variables. As evident from Cumulative Frequency and Percent, the data is divided into tertiles with first cut at 33.3 percent approx. and the second cut at 66.6 percent approx.
- The Categories are displayed as characters due to custom format being used that denies mapping between 0,1,2 ranks and ‘Low’, ‘Med’, ‘High categories as:

  0 -> Low
  1 -> Med
  2 -> High
- The variables have numeric values while the format defines the printing of these variables as character values for easy interpretation.

Relationship Between Derived Categorical Variables:

The relationship between the categorical variables is based on frequency rather than values. Thus, methods used for Numerical Correlations can not be used directly for describing the associations in categorical variables.

In this analysis, the following methods are used for examining the relationships between the variables:

- Graphical: Stacked bar chart
- Descriptive statistics: Cross Tables or 2x2 Contingency Tables
- Hypotheses testing: Chi-square tests to test if two categorical variables are independent
- Metric to measure the strength of the relation:
  - Cramer’s V

- Kendall's Tau
- Spherman's Rho

Contingency Table

A contingency table displays how two categorical variables are related in a table with how many values fall in each combination of categories. The categories of one variable define the rows and categories of the other variable define the columns of the table.

Chi-Square Tests

The statistical significance of associations obtained using contingency tables is obtained using Chi-Square Tests of Independence. The strength of the relationship between variables can also be obtained from the statistics obtained from Chi-Square Tests. Independence of two categorical variables means that knowing the outcome of one variable does not give you more information about the outcome of the other variable, and vice versa.

In this analysis, the chi-square test to determine the association between the categorical variables is used. For this analysis the degree of freedom (DF) is 4 since there are 3 rows and 3 columns in the contingency tables above and as per formula (r-1)*(c-1), the DF turns out to be 4. As can be seen from Decision Point Table, the DF of 4 has a decision point or critical value of 9.49. The first step is to define the Null and Alternative Hypothesis.

Null Hypothesis H0: The two variables are independent of each other among all subjects in population.
Alternate Hypothesis H1: The two variables are related to each other.
Interpretation:
If p-value is less than 0.05, we reject the null hypothesis otherwise we accept the null hypothesis. Also, if Chi-Squared Value is strictly larger than DP, then we have enough statistical evidence to reject the H0; and conclude that the two variables are not significantly independent to each other among all subjects in the population. Otherwise, we do not have enough statistical evidence to reject the H0; and conclude that X and Y are significantly independent to each other among all subjects in the population.

Chi-square Decision Points for Various Degrees of Freedom

df	dec.pt.	df	dec. pt.	df	dec. pt.
1	3.84	14	23.68	26	38.89
2	5.99	15	25.00	27	40.11
3	7.81	16	26.30	28	41.34
4	9.49	17	27.59	29	42.56
5	11.07	18	28.87	30	43.77
6	12.59	19	30.14	40	55.80
7	14.07	20	31.41	50	67.50
8	15.51	21	32.67	60	79.10
9	16.92	22	33.92	70	90.50
10	18.30	23	35.17	80	102.00
11	19.68	24	36.42	90	113.00
12	21.03	25	37.65	100	124.30
13	22.36				

Views

The Views on a YouTube video are an important parameter to describe the performance of video as well as the channel. The relationship of views with other prominent variables is analysed and described below:

Contingency Table Analysis:

Frequency
Expected
Deviation

Table of Views by Average_view_duration

Views(Views)	Average_view_duration(Average View Duration) Low	Med	High	Total
Low	45 23.78 21.22	15 24.44 -9.44	12 23.78 -11.78	72
Med	21 24.11 -3.11	26 24.78 1.2202	26 24.11 1.8899	73
High	6 24.11 -18.11	33 24.78 8.2202	34 24.11 9.8899	73
Total	72	74	72	218

Table of Views by Comments_added

Views(Views)	Comments_added(Comments Added) Low	Med	High	Total
Low	37 23.78 13.22	20 24.11 -4.11	15 24.11 -9.11	72
Med	28 24.11 3.8899	17 24.445 -7.445	28 24.445 3.555	73
High	7 24.11 -17.11	36 24.445 11.555	30 24.445 5.555	73
Total	72	73	73	218

Table of Views by Dislikes

Views(Views)	Dislikes(Dislikes) Low	Med	High	Total
Low	66 23.78 42.22	4 24.11 -20.11	2 24.11 -22.11	72
Med	6 24.11 -18.11	62 24.445 37.555	5 24.445 -19.44	73
High	0 24.11 -24.11	7 24.445 -17.44	66 24.445 41.555	73
Total	72	73	73	218

Table of Views by Impressions

Views(Views)	Impressions(Impressions) Low	Med	High	Total
Low	68 23.78 44.22	4 24.11 -20.11	0 24.11 -24.11	72
Med	4 24.11 -20.11	66 24.445 41.555	3 24.445 -21.44	73
High	0 24.11 -24.11	3 24.445 -21.44	70 24.445 45.555	73
Total	72	73	73	218

Table of Views by Likes

Views(Views)	Likes(Likes) Low	Med	High	Total
Low	61 23.78 37.22	10 24.11 -14.11	1 24.11 -23.11	72
Med	11 24.11 -13.11	51 24.445 26.555	11 24.445 -13.44	73
High	0 24.11 -24.11	12 24.445 -12.44	61 24.445 36.555	73
Total	72	73	73	218

Table of Views by Shares

Views(Views)	Shares(Shares) Low	Med	High	Total
Low	65 23.78 41.22	7 24.11 -17.11	0 24.11 -24.11	72
Med	7 24.11 -17.11	60 24.445 35.555	6 24.445 -18.44	73
High	0 24.11 -24.11	6 24.445 -18.44	67 24.445 42.555	73
Total	72	73	73	218

Table of Views by Subscribers

Views(Views)	Subscribers(Subscribers) Low	Med	High	Total
Low	64 23.78 40.22	8 24.11 -16.11	0 24.11 -24.11	72
Med	8 24.11 -16.11	56 24.445 31.555	9 24.445 -15.44	73
High	0 24.11 -24.11	9 24.445 -15.44	64 24.445 39.555	73
Total	72	73	73	218

Table of Views by Watch_time__hours_

Views(Views)	Watch_time__hours_(Watch Time Hours) Low	Med	High	Total
Low	67 23.78 43.22	5 24.11 -19.11	0 24.11 -24.11	72
Med	5 24.11 -19.11	62 24.445 37.555	6 24.445 -18.44	73
High	0 24.11 -24.11	6 24.445 -18.44	67 24.445 42.555	73
Total	72	73	73	218

The Contingency Tables above describe relations of Views with other variables. The following observations can be made:

- For Low and High values of Views, a relationship exists with Average View duration, Comments Added since the deviations from expected value are quite high while for Medium range of Views, the distribution is almost similar.
- The relationship of Views with all other variables i.e, Dislikes, Impressions, Likes, Shares, Subscribers and Watch Time (in Hours) is prominently visible since the values in tables differs largely from the expected values.

Chi Square Tests and Other Analysis:

Statistics for Table of Views by Average_view_duration

Statistic	DF	Value	Prob
Chi-Square	4	49.4142	<.0001
Likelihood Ratio Chi-Square	4	52.5583	<.0001
Mantel-Haenszel Chi-Square	1	38.6725	<.0001
Phi Coefficient		0.4761	
Contingency Coefficient		0.4299	
Cramer's V		0.3367	

Statistic	Value	ASE
Gamma	0.5325	0.0682
Kendall's Tau-b	0.3749	0.0524
Stuart's Tau-c	0.3749	0.0524
Somers' D C|R	0.3749	0.0524
Somers' D R|C	0.3749	0.0524
Pearson Correlation	0.4222	0.0567
Spearman Correlation	0.4220	0.0579

Statistics for Table of Views by Comments_added

Statistic	DF	Value	Prob
Chi-Square	4	33.7715	<.0001
Likelihood Ratio Chi-Square	4	37.4745	<.0001
Mantel-Haenszel Chi-Square	1	20.8972	<.0001
Phi Coefficient		0.3936	
Contingency Coefficient		0.3662	
Cramer's V		0.2783	

Statistic	Value	ASE
Gamma	0.3904	0.0727
Kendall's Tau-b	0.2703	0.0523
Stuart's Tau-c	0.2703	0.0523
Somers' D C|R	0.2703	0.0523
Somers' D R|C	0.2703	0.0523
Pearson Correlation	0.3103	0.0587
Spearman Correlation	0.3100	0.0596

Statistics for Table of Views by Dislikes

Statistic	DF	Value	Prob
Chi-Square	4	305.9780	<.0001
Likelihood Ratio Chi-Square	4	306.8664	<.0001
Mantel-Haenszel Chi-Square	1	174.4244	<.0001
Phi Coefficient		1.1847	
Contingency Coefficient		0.7642	
Cramer's V		0.8377	

Statistic	Value	ASE
Gamma	0.9707	0.0154
Kendall's Tau-b	0.8740	0.0272
Stuart's Tau-c	0.8740	0.0272
Somers' D C|R	0.8740	0.0272
Somers' D R|C	0.8740	0.0271
Pearson Correlation	0.8965	0.0249
Spearman Correlation	0.8965	0.0250

Statistics for Table of Views by Impressions

Statistic	DF	Value	Prob
Chi-Square	4	357.1609	<.0001
Likelihood Ratio Chi-Square	4	367.3726	<.0001
Mantel-Haenszel Chi-Square	1	196.5534	<.0001
Phi Coefficient		1.2800	
Contingency Coefficient		0.7880	
Cramer's V		0.9051	

Statistic	Value	ASE
Gamma	0.9966	0.0018
Kendall's Tau-b	0.9365	0.0164
Stuart's Tau-c	0.9365	0.0165
Somers' D C|R	0.9365	0.0164
Somers' D R|C	0.9365	0.0164
Pearson Correlation	0.9517	0.0129
Spearman Correlation	0.9518	0.0129

Statistics for Table of Views by Likes

Statistic	DF	Value	Prob
Chi-Square	4	217.1476	<.0001
Likelihood Ratio Chi-Square	4	225.6280	<.0001
Mantel-Haenszel Chi-Square	1	151.1085	<.0001
Phi Coefficient		0.9980	
Contingency Coefficient		0.7064	
Cramer's V		0.7057	

Statistic	Value	ASE
Gamma	0.9521	0.0157
Kendall's Tau-b	0.7924	0.0290
Stuart's Tau-c	0.7924	0.0291
Somers' D C|R	0.7924	0.0291
Somers' D R|C	0.7924	0.0290
Pearson Correlation	0.8345	0.0261
Spearman Correlation	0.8344	0.0262

Statistics for Table of Views by Shares

Statistic	DF	Value	Prob
Chi-Square	4	297.5885	<.0001
Likelihood Ratio Chi-Square	4	305.2399	<.0001
Mantel-Haenszel Chi-Square	1	179.8328	<.0001
Phi Coefficient		1.1684	
Contingency Coefficient		0.7597	
Cramer's V		0.8262	

Statistic	Value	ASE
Gamma	0.9880	0.0047
Kendall's Tau-b	0.8833	0.0215
Stuart's Tau-c	0.8833	0.0216
Somers' D C|R	0.8833	0.0215
Somers' D R|C	0.8833	0.0215
Pearson Correlation	0.9103	0.0175
Spearman Correlation	0.9104	0.0175

Statistics for Table of Views by Subscribers

Statistic	DF	Value	Prob
Chi-Square	4	262.0314	<.0001
Likelihood Ratio Chi-Square	4	271.4874	<.0001
Mantel-Haenszel Chi-Square	1	169.0986	<.0001
Phi Coefficient		1.0963	
Contingency Coefficient		0.7388	
Cramer's V		0.7752	

Statistic	Value	ASE
Gamma	0.9789	0.0071
Kendall's Tau-b	0.8484	0.0240
Stuart's Tau-c	0.8484	0.0241
Somers' D C|R	0.8484	0.0240
Somers' D R|C	0.8484	0.0240
Pearson Correlation	0.8828	0.0200
Spearman Correlation	0.8827	0.0200

Statistics for Table of Views by Watch_time__hours_

Statistic	DF	Value	Prob
Chi-Square	4	316.6811	<.0001
Likelihood Ratio Chi-Square	4	324.1450	<.0001
Mantel-Haenszel Chi-Square	1	185.3237	<.0001
Phi Coefficient		1.2053	
Contingency Coefficient		0.7696	
Cramer's V		0.8523	

Statistic	Value	ASE
Gamma	0.9915	0.0036
Kendall's Tau-b	0.9008	0.0201
Stuart's Tau-c	0.9008	0.0202
Somers' D C|R	0.9008	0.0201
Somers' D R|C	0.9008	0.0201
Pearson Correlation	0.9241	0.0161
Spearman Correlation	0.9241	0.0162

The above statistical tables describe various measures of Chi-Sqaure and other statistics for correlation between the variables. The following observations are made:

- The Chi-Square values are significanlty higher than the decision point/ critical value (for DF of 4 from descision point table) for majority of variables except Average_View_Duration and Comments_Added which are quite low in comparion to other variables. Also, the p-value for all these Chi-Square values is much less than 0.05. Thus, it can be inferenced that based on Chi-Square Test, the relationship between Views and Dislikes, Impressions, Likes, Shares, Subscribers and Watch Time Hours is statistically Significant.
- Similar interpetation can be made from Crammer's V value as it is close to zero for Average View Duration and Comments Added but close to one for other variables.
- The Kendall's Tau-b and Spearman Correlation coeffecient also results in similar interpretation. While the values are less tha 0.43 for Average_View_Duration and Comments_Added describing weak relationships, the values are between 0.8 and 1 for other variables signifying very strong relationship with Views.

Average Percentage Viewed

The Average Percentage Viewed on a YouTube video is an important parameter to describe the interest of the user in the content. Lower interest results in early closure of video resulting in less amount of video viewed. The relationship of Average Percentage Viewed with other prominent variables is analysed and described below

Contingency Table Analysis:

Frequency
Expected
Deviation

Table of Average_percentage_viewed____ by Clicks_per_end_screen_element_sh				
	Clicks_per_end_screen_element_sh(Clicks/End Screen Element Shown)			
Average_percentage_viewed____(Average Percentage Viewed)	Low	Med	High	Total
Low	34 23.78 10.22	24 24.11 -0.11	14 24.11 -10.11	72
Med	21 24.11 -3.11	25 24.445 0.555	27 24.445 2.555	73
High	17 24.11 -7.11	24 24.445 -0.445	32 24.445 7.555	73
Total	72	73	73	218

Table of Average_percentage_viewed____ by Impressions_click_through_rate__				
	Impressions_click_through_rate__(Impressions CTR)			
Average_percentage_viewed____(Average Percentage Viewed)	Low	Med	High	Total
Low	34 24.11 9.8899	14 23.78 -9.78	24 24.11 -0.11	72
Med	15 24.445 -9.445	30 24.11 5.8899	28 24.445 3.555	73
High	24 24.445 -0.445	28 24.11 3.8899	21 24.445 -3.445	73
Total	73	72	73	218

The Contingency Tables above describe relations of Average Percentage Viewed with variables Clicks_Per_End_Screen_Element_Shown_ and Impressions_click_through_rate. The following observations can be made:

- For Low and High values of Average_Percentage_Viewed and corresponding Low and High values Clicks_Per_End_Screen_Element, the deviations from expected value are a little higher than other cases that signify that these categories might be related.
- For Low and Med values of Average_Percentage_Viewed and corresponding Low and Med values Clicks_Per_End_Screen_Element, the deviations from expected value are a little higher than other cases that signify that these categories might be related.

Chi Square Tests and Other Analysis:

Statistics for Table of Average_percentage_viewed____ by Clicks_per_end_screen_element_sh

Statistic	DF	Value	Prob
Chi-Square	4	13.7532	0.0081
Likelihood Ratio Chi-Square	4	14.0363	0.0072
Mantel-Haenszel Chi-Square	1	12.6408	0.0004
Phi Coefficient		0.2512	
Contingency Coefficient		0.2436	
Cramer's V		0.1776	

Statistic	Value	ASE
Gamma	0.3172	0.0827
Kendall's Tau-b	0.2148	0.0577
Stuart's Tau-c	0.2148	0.0577
Somers' D C|R	0.2148	0.0577
Somers' D R|C	0.2148	0.0577
Pearson Correlation	0.2414	0.0644
Spearman Correlation	0.2412	0.0645

Statistics for Table of Average_percentage_viewed____ by Impressions_click_through_rate__

Statistic	DF	Value	Prob
Chi-Square	4	14.8058	0.0051
Likelihood Ratio Chi-Square	4	15.5012	0.0038
Mantel-Haenszel Chi-Square	1	0.5023	0.4785
Phi Coefficient		0.2606	
Contingency Coefficient		0.2522	
Cramer's V		0.1843	

Statistic	Value	ASE
Gamma	0.0595	0.0937
Kendall's Tau-b	0.0403	0.0636
Stuart's Tau-c	0.0403	0.0636
Somers' D C|R	0.0403	0.0636
Somers' D R|C	0.0403	0.0636
Pearson Correlation	0.0481	0.0697
Spearman Correlation	0.0478	0.0707

The above statistical tables describe various measures of Chi-Sqaure and other statistics for correlation between the variables. The following observations are made:

- The Chi-Square values are quite low and near to the decision point value of 9.49. Also, the p-value for all these Chi-Square values is less than 0.05. Thus, it can be inferenced that based on Chi-Square Test, the relationship between Average_Percentage_Viewed and Clicks_per_end_Screen_element_shown, Impressions_CTR is slightly statistically significant.
- Similar interpetation can be made from Crammer's V value as it is close to zero signifying very weak relationship.
- The Kendall's Tau-b and Spearman Correlation coeffecient also results in similar interpretation. The values are very close to zero for both relations signifying very low or no corelation.

Clicks Per End Screen Element Shown

The Clicks per end screen elements on a YouTube video is an the number of clicks after a video has ended. These clicks can lead user to become subscriber or watch some other content of the same channel. The relationship of Clicks per end screen element shown with Impressions CTR is described below

Contingency Table Analysis:

Frequency
Expected
Deviation

Table of Clicks_per_end_screen_element_sh by Impressions_click_through_rate__

Clicks_per_end_screen_element_sh(Clicks/End Screen Element Shown)	Impressions_click_through_rate__(Impressions CTR)			
	Low	Med	High	Total
Low	34 24.11 9.8899	13 23.78 -10.78	25 24.11 0.8899	72
Med	23 24.445 -1.445	27 24.11 2.8899	23 24.445 -1.445	73
High	16 24.445 -8.445	32 24.11 7.8899	25 24.445 0.555	73
Total	73	72	73	218

The Contingency Tables above describe relations of Clicks_Per_End_Screen_Element_Shown_ with Impressions_click_through_rate. The following observations can be made:

- For Low and High values of Clicks_Per_End_Screen_Element_Shown_ and corresponding Low and Med values Impressions_click_through_rate, the deviations from expected value are a little higher than other cases that signify that these categories might be related.

Chi Square Tests and Other Analysis:

Statistics for Table of Clicks_per_end_screen_element_sh by Impressions_click_through_rate__

Statistic	DF	Value	Prob
Chi-Square	4	15.0056	0.0047
Likelihood Ratio Chi-Square	4	15.6711	0.0035
Mantel-Haenszel Chi-Square	1	3.3212	0.0684
Phi Coefficient		0.2624	
Contingency Coefficient		0.2538	
Cramer's V		0.1855	

Statistic	Value	ASE
Gamma	0.1634	0.0902
Kendall's Tau-b	0.1108	0.0616
Stuart's Tau-c	0.1108	0.0616
Somers' D C|R	0.1108	0.0616
Somers' D R|C	0.1108	0.0616
Pearson Correlation	0.1237	0.0679
Spearman Correlation	0.1237	0.0689

The above statistical tables describe various measures of Chi-Sqaure and other statistics for correlation between the variables. The following observations are made:

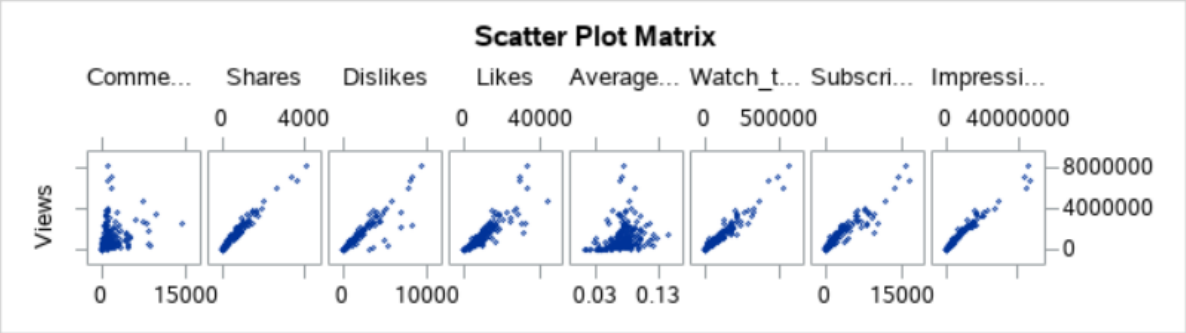
- The Chi-Square values are quite low and near to the decision point value of 9.49. Also, the p-value for all these Chi-Square values is less than 0.05. Thus, it can be inferenced that based on Chi-Square Test, the relationship between Clicks_per_end_Screen_element_shown and Impressions_CTR is slightly statistically significant.
- Similar interpetation can be made from Crammer's V value as it is close to zero signifying very weak relationship.
- The Kendall's Tau-b and Spearman Correlation coeffecient also results in similar interpretation. The values are very close to zero for both relations signifying very low or no corelation.

Comparing Relationship between Numerical and Categorical Variables:

In this section, a comparison has been made between the correlation statistics obtained from numerical dataset and the correlation obtained above using various test on categorical dataset. This comparison is intended to affirm that these results are similar and verify that no information loss has occurred due to conversion to categorical variables.

Similar to previous analysis, Views variable is correlated with other variables.

1 With Variables:	Views
8 Variables:	Comments_added Shares Dislikes Likes Average_view_duration Watch_time__hours_ Subscribers Impressions



The following observations can be made from below tables and above correlation matrix:

- There is monotonous relationship between Views and other variables Comments_Added, Shares, Dislikes, Likes, Average_View_Duration, Watch_time_hours_, Subscribers and Impressions.
- The relationship is relatively stronger for lower values of variables as compared to higher ones. Also, the relation is positive i.e, increase in one variable results in increase in another.
- The tables below quantify the relationships. The p-value is very less than 0.05 for all variables showing that the results obtained are statistically significant. This is inline with the results obtained from Chi-Sqaure statistics for categorical variables.
- The Person's Coeffecint is close to zero for relation between Views and Comments_Added, Average_View_Duration while is it very close to one for others signifying a weak relationship of Comments_Added, Average_View_Duration with Views and a very strong relationship with other variables.
- The Spearman and Kandall Tau coeffecients obtained using the numerical data also show similar values when compared to one obtained from categorical Data.
- Thus, it can be inferred that the relationships and their sterngth remained same even after binning the numerical data to categorical data.

Pearson Correlation Coefficients, N = 218 Prob > |r| under H0: Rho=0	Views
Comments_added	0.23699 0.0004
Shares	0.98193 <.0001
Dislikes	0.90270 <.0001
Likes	0.88450 <.0001
Average_view_duration	0.27876 <.0001
Watch_time__hours_	0.97017 <.0001
Subscribers	0.94093 <.0001
Impressions	0.98155 <.0001

Spearman Correlation Coefficients, N = 218 Prob > |r| under H0: Rho=0	Views
Comments_added	0.42566 <.0001
Shares	0.98279 <.0001
Dislikes	0.94090 <.0001
Likes	0.94696 <.0001
Average_view_duration	0.46285 <.0001
Watch_time__hours_	0.97839 <.0001
Subscribers	0.96515 <.0001
Impressions	0.99092 <.0001

Kendall Tau b Correlation Coefficients, N = 218 Prob > |tau| under H0: Tau=0	Views
Comments_added	0.30864 <.0001
Shares	0.89037 <.0001
Dislikes	0.86633 <.0001
Likes	0.81422 <.0001
Average_view_duration	0.32774 <.0001
Watch_time__hours_	0.88534 <.0001
Subscribers	0.84981 <.0001
Impressions	0.92305 <.0001

Determining Predictors and Response Variable:

Based on the analysis and interpretation developed, the following variables are selected:

Response Variable: Views

The Views is the most important metric for videos published on YouTube. It is dependent on a lot factors that maintain the quality of channel and keep the users attracted.

Predictors:

As seen in the categorical data analysis as well as comparison with the Numerical data, the following variables exhibit strong correlations with the views. Thus, these variables can be select as the predictors of the response variable Views:

- Shares
- Likes
- Watch_Time_Hours_
- Subscribers
- Impressions

These variables are further analysed by creating Linear Regression Models to decide predictors based on the statistics obtained.

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	1	3.301266E14	3.301266E14	5814.76	<.0001
Error	216	1.226317E13	56773923333		
Corrected Total	217	3.423898E14			

Root MSE	238273	R-Square	0.9642
Dependent Mean	1126416	Adj R-Sq	0.9640
Coeff Var	21.15317		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	146617	20628	7.11	<.0001	0
Shares	1	2100.73285	27.54895	76.25	<.0001	0.98193

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	1	2.678652E14	2.678652E14	776.37	<.0001
Error	216	7.452452E13	3.450209E11		
Corrected Total	217	3.423898E14			

Root MSE	587385	R-Square	0.7823
Dependent Mean	1126416	Adj R-Sq	0.7813
Coeff Var	52.14633		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-109153	59574	-1.83	0.0683	0
Likes	1	150.50941	5.40167	27.86	<.0001	0.88450

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	1	3.222671E14	3.222671E14	3459.27	<.0001
Error	216	2.012265E13	93160410476		
Corrected Total	217	3.423898E14			

Root MSE	305222	R-Square	0.9412
Dependent Mean	1126416	Adj R-Sq	0.9410
Coeff Var	27.09672		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	52989	27576	1.92	0.0560	0
Watch_time__hours_	1	12.92253	0.21971	58.82	<.0001	0.97017

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	1	3.03133E14	3.03133E14	1667.91	<.0001
Error	216	3.925678E13	1.817444E11		
Corrected Total	217	3.423898E14			

Root MSE	426315	R-Square	0.8853
Dependent Mean	1126416	Adj R-Sq	0.8848
Coeff Var	37.84700		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	264780	35760	7.40	<.0001	0
Subscribers	1	411.67242	10.08012	40.84	<.0001	0.94093

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	1	3.298712E14	3.298712E14	5691.73	<.0001
Error	216	1.251855E13	57956271833		
Corrected Total	217	3.423898E14			

Root MSE	240741	R-Square	0.9634
Dependent Mean	1126416	Adj R-Sq	0.9633
Coeff Var	21.37230		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	148798	20827	7.14	<.0001	0
Impressions	1	0.16392	0.00217	75.44	<.0001	0.98155

The following observations can be made from these statistics:

- All the predictors are statistically significant based on p-value. Also, the models are also significant based on p-value of F-statistics.
- The most dominant predictors based on Parameter estimates on slopes are Shares, Subscribers and Likes.
- The other variables have low slope estimates signifying that these variables have lower effect on the response variable as compared to the above mentioned three variables.

Thus, on the basis of the estimated values in the analysis above, it can be decided that:

- Dependent Variable: Views
- Predictors: Shares, Likes and Subscribers

Linear Regression and Prediction:

The analysis done in previous section provided three most significant predictors of response variable Views. These are Subscribers, Likes and Shares. In this section, a Linear Regression Model is developed using these predictors to predict the values of Views.

The following steps are taken to perform Predication of Views from Likes, Shares and Subscribers:

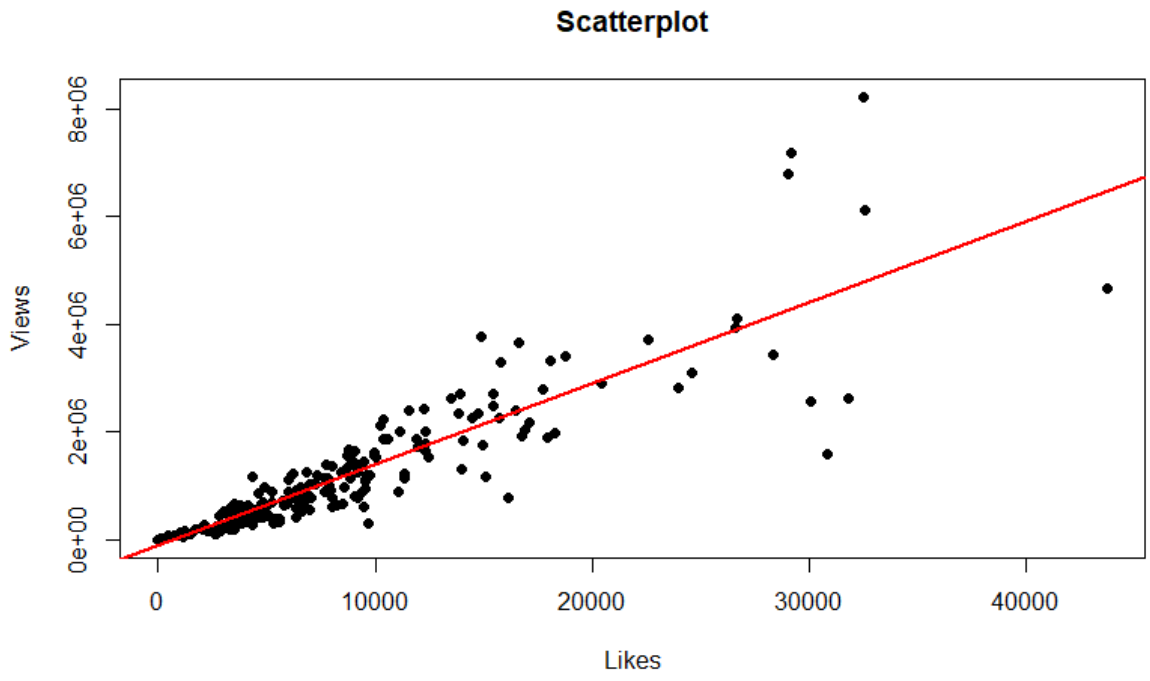
- A Linear Model is constructed using Views as dependent variable and Likes as independent variable. Further, a plot between these variables along with the regression line is drawn to display the linear relationship. Finally, the value of Views is predicted from model using predict() method in R.
- The process is repeated for Shares and Subscribers to predict Views based on these variables.

Likes

The following linear model is constructed:

Regression Equation: **Views = -109153 + 150.5*Likes**

Scatter Plot with Regression Line:



Prediction:

The following values are predicted using the model:

Likes	Views
25000	3653582
26000	3804091

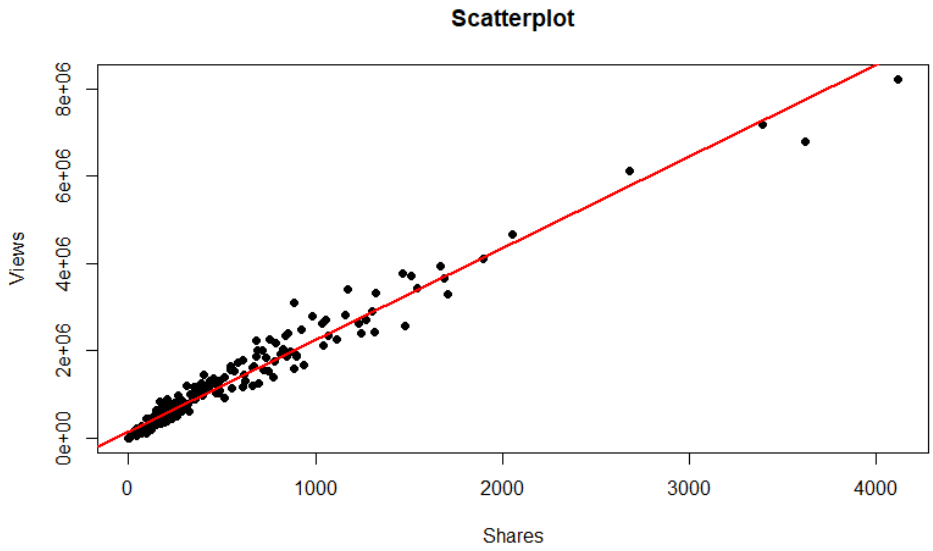
As evident from the predictions as well as scatterplot, the number Views of the video increase with increase in Likes. An increase of 1000 likes resulted in almost 150000 more views on video.

Shares

The following linear model is constructed:

Regression Equation: **Views = 146617 + 2100.73*Shares**

Scatter Plot with Regression Line:



Prediction:

The following values are predicted using the model:

Shares	Views
2200	4768230
2800	6028669

As evident from the predictions as well as scatterplot, the number Views of the video increase with increase in Shares. An increase of 400 likes resulted in almost 1200000 more

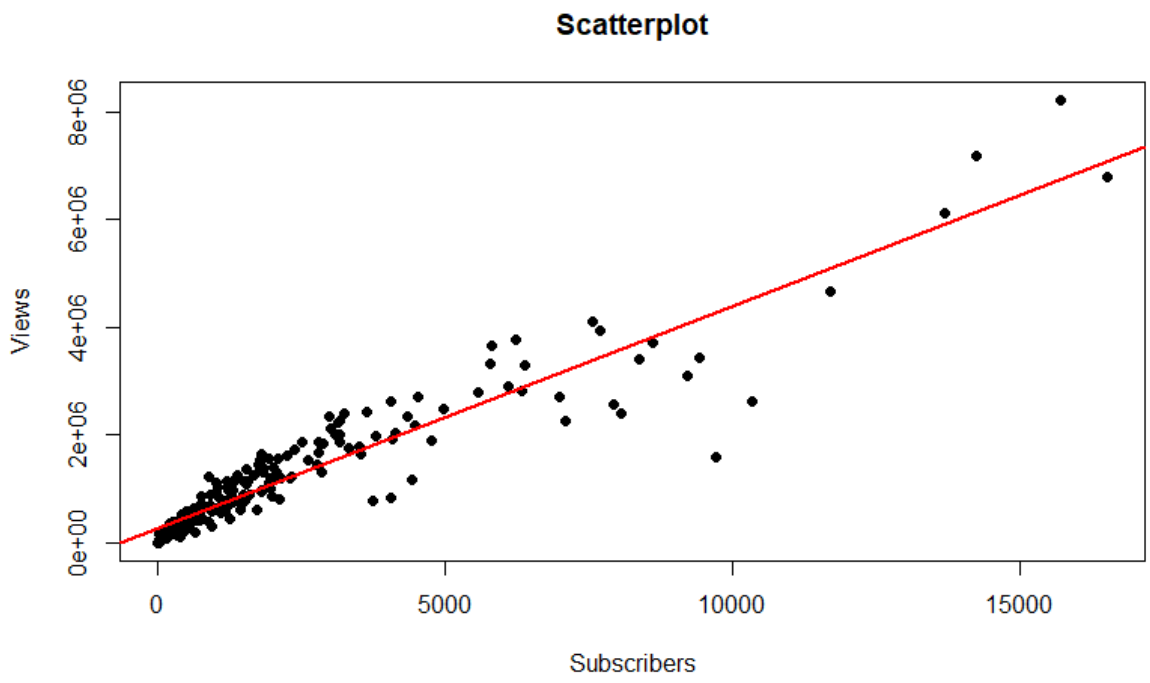
views on video. This is a significant increase in the views with relatively lower increase in shares.

Subscribers

The following linear model is constructed:

Regression Equation: **Views = 264780.44 + 411.67*Subscribers**

Scatter Plot with Regression Line:



Prediction:

The following values are predicted using the model:

Subscribers	Views
7500	3352324
8500	3763996

As evident from the predictions as well as scatterplot, the number Views of the video increase with increase in number of Subscribers added. An increase of 1000 subscribers resulted in almost 400000 more views on video.

Recommendations:

Based on the analysis, tests an interpretation obtained above, the following recommendations are suggested to the Administrator of the YouTube channel:

- The most prominent factor of increasing Views on the video is the number of times it is shared on the platform. Thus, an effort must be made to share the video extensively. Also, a healthy practice is to request your viewers to share your content across.
- The number of Subscribers added per video is the next parameter that enhances views on the video. In order to achieve more subscribers per video, it is recommended to create more engaging content that is liked by the audience. A viewer turns into a subscribers if the video manages to keep the interest till the end and excites viewer to watch more related content. Thus, it is recommended to craft the content targeted towards the interests of audience in order to achieve more subscribers and views. The same concept goes for number of likes on the video.

Summary:

In this report, analysis of YouTube Video data is performed. The report is summarized below

- The Numerical data present in dataset is first Categorized into 3 categories namely 'Low', 'Med' and 'High' based on the value using Adaptive Quantile-Based binning technique.
- The relationship between the categorical variables of derived dataset is analyzed by performing various tests as mentioned below
    - Contingency Tables: to determine independence using frequency distributions between pair of variables
    - Chi-Square Tests statistics: To obtain the statistical significance of relationships
    - Kendall's Tau-b and Spearman's Correlation Tests: to obtain strength of correlations.
- A comparison between the relationships obtained from derived categorical dataset and numerical data set is performed by evaluating Pearson's Correlation between numerical variables. Similar results were observed for numerical and categorical datasets.
- Based on results of correlation tests and strength of correlation, 'Views' is selected as dependent variable and Shares, Likes and Subscribers as Predictors for performing Linear Regression.
- Linear Regression is performed to predict values of Views based on selected predictors.
- The YouTube Administrator is suggested with useful recommendations increasing the number of Views on the Videos.