Twitch Data Analysis Report

Alina Baciu

6/8/23

Statistics for Data Scientists

Table of Contents

1. Context

The client of this assignment received negative feedback from his academia superiors concerning his performance throughout the year. As a consequence, the client decided to change his career and become a content creator on Twitch, a live video streaming platform with a variety of broadcasts such as related to gaming, cooking, sports and music.

2. Problem

The client faces multiple questions concerning the strategic approach he needs to adopt due to its lack of experience and knowledge in the online content production field. The data set that serves as the starting point for this analysis can be accessed on the Kaggle platform through the following link: https://www.kaggle.com/datasets/aayushmishra1512/twitchdata.

The data set presents different performance metrics and characteristics regarding the top 1000 Twitch streamers from the past year, such as followers, viewership, stream duration, language, maturity classification of the streamers' content, and their Twitch partnership status.

The given inquiries are the following:

I. Should I focus on mature content for a grown-up audience? What are the consequences if I decide to do so in terms of audience reactions? Is the effect constant across all different ways one can think about "performance".
II. Does this choice lower or increase my chance of becoming a Twitch partner?
III. As you know, I am sluggish: Is the effect of the minutes I stream larger or smaller in mature content?

3. Goal

The goal of the assignment is to conduct a data analytics case study on the Twitch streamers data and use the gained insights to make a recommendation to support the decision-making process of the client and fulfill his professional goal.

Resolving the given questions is of the utmost importance in formulating an effective strategic approach to achieve success as a content creator. While the current data set represents a good starting point for the analysis, it is crucial for the student to also assess potential limitations of the chosen approach. Other strategies of new data acquisition are also expected to be critically evaluated to address these limitations and therefore to enhance the analysis.

4. Data Understanding

4.1 Data Collection

The CSV file was imported as a data set in the R Studio Quarto environment using a relevant built-in function. The "knitr" package was used to generate a formatted table to improve the readability of the presented data.

```
library(knitr)

data <- read.csv("twitchdata-update.csv")
knitr::kable(data[1:5, 1:5])
```

Channel	Watch.time.Minutes.	Stream.time.minutes.	Peak.viewers	Average.viewers
xQcOW	6196161750	215250	222720	27716
summit1g	6091677300	211845	310998	25610
Gaules	5644590915	515280	387315	10976
ESL_CSGO	3970318140	517740	300575	7714
Tfue	3671000070	123660	285644	29602

Table 1. Twitch data sample

4.2 Data Description

Three tasks were performed on the data set to discover initial insights: volumetric analysis, data type classification and target analysis. Concerning the volumetric analysis, the Twitch data set consists of 11 attributes and 1000 observations, each instance representing the channel characteristics of a top streamer. The table above (see Table 1) depicts a sample of this data including the first 5 fields and observations. The "Channel" variable is the primary key of the table as it is used to uniquely identify each streamer.

The data type classification from a statistical point of view can be found in Table 2 and its role is to facilitate the creation of data visualizations.

Variable	Data Type
Channel	Qualitative - Nominal
Watch time (in minutes)	Quantitative - Discrete, Ratio
Streaming time (in minutes)	Quantitative - Discrete, Ratio

Variable	Data Type
Peak viewers	Quantitative - Discrete, Ratio
Average viewers	Quantitative - Discrete, Ratio
Followers	Quantitative - Discrete, Ratio
Followers gained	Quantitative - Discrete, Ratio
Views gained	Quantitative - Discrete, Ratio
Partnered	Qualitative - Nominal
Mature	Qualitative - Nominal
Language	Qualitative - Nominal

Table 2. Statistical Data Type for each column

Concerning the target analysis, the provided inquiries will be assessed to determine the dependent/independent variables.

The first point focuses on the decision to focus on mature content when opening a Twitch channel. The dependent variable is represented by the audience reactions (viewership - Average viewers, Peak viewers, Views Gained, Watch time) and followers activity - (Followers Gained, Followers)) while the independent variable is mainly Mature (but Language and Partnered will also be checked). In this way, it will be noticed whether the presence or the absence of mature content influences the engagement and the behavior of the audience. Moreover, analyzing the relationship between the Mature column and audience reactions will determine whether there exists a significant pattern, a constant effect among all performance metrics.

The second point implies whether the decision to focus on mature content lowers or increases the chances of becoming a Twitch partner. The dependent variable is the Partnered column while the independent one is mainly Mature (but Language will also be checked). In this manner, it can be examined whether the presence or absence of mature content determines a significant relationship with becoming a partner.

Ultimately, regarding the third point, the client mentioned that he is sluggish, meaning that he tends to be slow-moving when talking. Therefore, he is interested in whether the presence or absence of mature content (Mature - independent variable; but Language and Partnered will also be checked) leads to an increase or decrease of the streamed minutes (Streaming time - dependent variable).

4.3 Exploratory Data Analysis

The student performed two tasks to understand the relationship between features: data analysis (univariate/bivariate, including correlation analysis) and data cleaning (checking missing values).

4.3.1 Data Analysis

4.3.1.1 Univariate Analysis

There is a majority of TRUE values for Partnered variable (97%) and a majority of FALSE values for Mature variable (77%). The channels are transmitted in 21 different languages, with a majority of English channels (48%).

Regarding the quantitative variables, the reader can observe that all features have a right-skewed distribution (see Figure 1 - Appendix).

These distributions suggest that there is a small number of channels that have high performance metric values. The median should be used in this scenario as measure of center as it is more representative than the mean and less affected by extreme values.

4.3.1.2 Bivariate Analysis

I. Mature variable vs Audience Reactions

As mentioned in the Data Description chapter, to answer the first point of the assignment, we need to analyze the relationship between Mature variable and the Audience Reactions.

- Viewership (Average viewers, Peak viewers, Views gained, Watch time Minutes)

As we deal with two different samples (mature/non-mature content channels), the difference in medians technique will be applied to determine which one has higher performance and by how much. After calculating the difference between the medians, the certainty of the results will be generated using the Bootstrapping technique, showing the estimation mean, standard deviation and also the 95% confidence interval.

```
suppressPackageStartupMessages({library(tidyverse)})
library(tibble)
library(dplyr)
mature = filter(data,data[['Mature']]=="True")
non_mature = filter(data,data[['Mature']]=="False")
sample_mature <- function(){return(slice_sample(mature,n=nrow(mature),
replace=TRUE)[["Average.viewers"]])}
sample_nonmature <- function(){ return(
slice_sample(non_mature,n=nrow(non_mature),replace=TRUE)
[["Average.viewers"]])}
bs_dist <- tibble( d_in_means=replicate(1000,median(sample_nonmature())
-median( sample_mature() ) ) )
# mean( bs_dist[["d_in_means"]] ) # sd( bs_dist[["d_in_means"]] )
```

```
# plot <- ggplot(bs_dist,aes(x="",y=d_in_means))
#+geom_quasirandom(col="gray")+
# stat_summary(col="red",fun.data=qrange(95),geom="errorbar")
#+xlab("") + coord_flip() # upper <- plot_data$ymax[1]
# plot_data <- ggplot_build(plot)$data[[2]] # lower <- plot_data$ymin[1]
```

	Difference	Estimation Mean and SE	Confidence Interval 95%
Avg viewers	485,5 (False)	499.9 +-173.1	168 - 859
Peak viewers	5.227,5 (False)	5187 +-1436.8	2504 - 8108
Views gained	1.713.678 (False)	1.842.259 +-370.274	1.232.977 - 2.626.370
Watch min	30.131.66 (False)	29.346.391 +-15.102.27	-18.322.981 - 56.709.776

Table 3. Difference in Medians. Viewership vs Mature

Based on the results, it appears that the non-mature content has a constant growth effect on all viewership metrics.

Concerning the magnitude of the impact, the effect appears to vary in strength for all variables (Appendix: Viewership vs Mature). There appears to be a wider CI for Watch Min, suggesting the uncertainty of its outcome in growth or decrease.

"Average viewers" value may indicate that the non-mature content has a better ability to attract the audience in a particular stream session. However, the higher difference value for "Watch Time (in minutes)" attribute may suggest that such a channel has a better ability to engage the audience for a longer time once they enter a stream, leading to a longer watch time for the overall channel. The higher difference for "Views Gained" can show that a channel generating non-mature content collects a considerably higher number of overall views (e.g. from live, archive streams) during a year, even though it may not attract a considerably higher number of users in a streaming session. Ultimately, the difference of the "Peak Viewers" variable may suggest that there are times when channels with non-mature content do manage to attract a larger number of viewers in a stream.

- Followers Activity (Followers, Followers Gained)

	Difference	Estimation Mean and SE	Confidence Interval 95%
Followers	58.431 (False)	59.662 +-22.859	18.773 - 102.559
Followers G.	38.037 (False)	38.181 +-9567	17.470 - 55.094

Table 4. Difference in Medians. Followers vs Mature

The results from Table 4 suggest a constant increase effect on followers activity metrics, varying in strength (Appendix: Followers Activity vs Mature). A channel that has non-mature content gains in a year approximately 38k more followers and has approximately 58k more overall followers. This followers increase may be mostly represented by the young users who tend to be more active on such platforms these days possibly due to their more free time and who watch this type of content as it aligns more with their entertainment preferences.

There appears to be a wider CI for Followers, suggesting the uncertainty of its outcome in growth. However, the significance of the hypotheses will be further analyzed in the Confirmatory Data Analysis Chapter.

II. Mature variable vs Partnered variable

To answer the second point of the assignment, we need to analyze the relationship between the Mature variable and Partnered. Figure 8 shows the partnership rates per each value of the Mature variable.

```
library(ggplot2)
suppressMessages(library(GGally))
data$Mature <- as.factor(data$Mature)
ggally_colbar(data, mapping = aes(x = Mature, y = Partnered), size = 2.0)
```

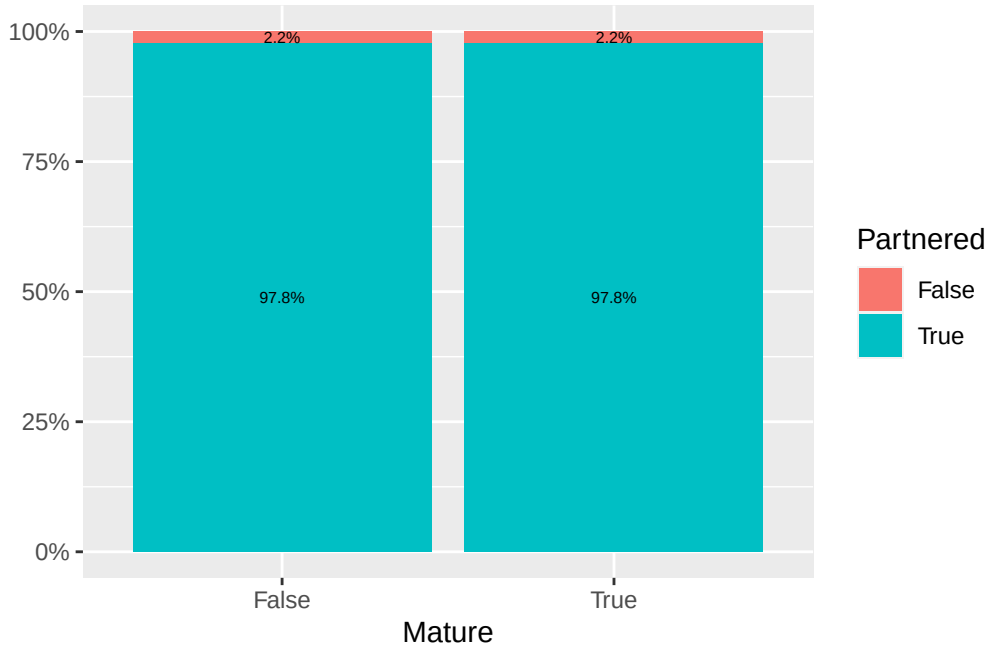

Figure 8. Mature vs Partnered

The visualization determines that the content nature does not indicate any impact on the chance of becoming a Twitch partner. The rates appear to be almost identical as the Partnered - True streamers represent 97.79% for the Non-mature type and 97.83% for the Mature category.

A channel with mature content appears to have a slightly higher partnership rate than non-mature though but the difference is really small and probably not notable. The ratio of proportions (97.83 % / 97.79 % = 1.0004) also indicates no visible effect.

Moreover,the Bootstrapping results support these findings, suggesting an even smaller estimation. Even though the EDA could not determine any patterns, further investigation will be done in the CDA.

```
library(dplyr)
library(tibble)
library(tidyverse)

M = filter(data,data[['Mature']]=="True")
NM = filter(data,data[['Mature']]=="False")
perc_part_true_mat <- sum(M$Partnered == "True") / nrow(M) * 100
perc_part_false_mat <- sum(M$Partnered == "False") / nrow(M) * 100
perc_part_true_nonmat <- sum(NM$Partnered == "True") / nrow(NM) * 100
perc_part_false_nonmat <- sum(NM$Partnered == "False") / nrow(NM) * 100
ratio_percentages <- perc_part_true_mat / perc_part_true_nonmat
```

Ratio of proportions	Estimation Mean and SE	Confidence Interval 95%
1.004	1.002 +-0.01	0.97 - 1.02

Table 5. Ratio of proportions

III. Mature variable vs Streaming time (in minutes) variable

The EDA outcome suggest that the effect of streaming time may be larger for mature content. The mature type has approximately 16k more streaming minutes than the non-mature category.

```
ggplot(data, aes(x = Mature, y = Stream.time.minutes., fill = Mature)) +
geom_bar(stat = 'summary', fun = 'median') +
stat_summary(geom = 'text', fun = 'median', aes(label = after_stat(y),
group = Mature), vjust = -0.1)
```

Figure 9. Mature vs Streaming Time Minutes

The estimation mean is 15.430 min with a SE of 5.718. The higher difference can show that a channel generating mature content may be involved in more complex discussions or debatable topics, leading to longer streaming duration. Further steps will be taken in CDA, to assess the significance of the pattern.

	Difference	Estimation Mean +-SE	Confidence Interval 95%
Streaming Min	16.342 (True)	15.430 + 5.718	4.124 - 25.847

Table 6. Difference. Streaming Time

IV. Correlation Analysis

When calculating the correlation scores between variables, we can identify three types of associations: numerical vs numerical, categorical vs categorical and categorical vs numerical. When it comes to generating the scores between Mature and the Audience Reactions metrics and also the scores between Mature and Streamed Time (in minutes) the correlation was not calculated as the Point biserial technique assumes that the numerical variables are normally distributed. The correlation score between the categorical variables was calculated using the

Cramer's V method (chi-squared statistic). The 0.000972 score between Partnered and Mature indicates a very weak association as observed below. The correlation between Partnered, Mature and Language (0.25 and 0.24) also suggest a weak relationship.

```
library(rcompanion)
cramer_v <- cramerV(data$Mature, data$Partnered)
print(cramer_v)
```

```
Cramer V
0.000972
```

Next, the association between the Audience Reaction variables was analyzed as the student was curious whether an increase in one attribute also implies an increase in the other one. To achieve this, the Pearson's correlation technique was used for the numerical vs numerical variables.

Figure 10 depicts the following strong associations (0.60 - 0.79) : Average viewers and Peak viewers (0.68); Followers and Watch time Minutes (0.62); Followers and Followers gained (0.71); In other words, an increase in average viewers corresponds to an increase in peak viewership; an increase in watch time corresponds to an increase in followers; an increase in followers gained in a year corresponds to an increase in followers;

```
suppressPackageStartupMessages(library(corrplot))
data_q <- data[,2:8]
corr <- cor(data_q)
corr
```

```
                    Watch.time.Minutes. Stream.time.minutes. Peak.viewers
Watch.time.Minutes.           1.0000000           0.15058790    0.5827966
Stream.time.minutes.          0.1505879           1.00000000   -0.1195403
Peak.viewers                  0.5827966          -0.11954029    1.0000000
Average.viewers               0.4761650          -0.24924779    0.6826373
Followers                     0.6202339          -0.09129851    0.5325293
Followers.gained              0.5146476          -0.15816479    0.4704147
Views.gained                  0.5298620           0.06437003    0.2980626
                    Average.viewers   Followers Followers.gained Views.gained
Watch.time.Minutes.       0.4761650  0.62023388        0.5146476   0.52986201
Stream.time.minutes.     -0.2492478 -0.09129851       -0.1581648   0.06437003
Peak.viewers              0.6826373  0.53252932        0.4704147   0.29806263
Average.viewers           1.0000000  0.42830322        0.4200974   0.25034887
Followers                 0.4283032  1.00000000        0.7156185   0.27646651
```

```
Followers.gained          0.4200974  0.71561846        1.0000000   0.24429687
Views.gained              0.2503489  0.27646651        0.2442969   1.00000000
```

Figure 10. Correlation Table

V. Data Cleaning

```
has_null_values <- colSums(is.na(data)) > 0
print(names(has_null_values[has_null_values]))
```

```
character(0)
```

4.4 Confirmatory Data Analysis

I. Mature variable vs Audience Reactions

1. Hypothesis Testing - Bootstrapping

The Exploratory Data Analysis results indicated that the non-mature content has a constant growth effect on all viewership and followers activity metrics. To check whether we are wrong about this statement, the null hypothesis of "no difference" was used (no difference in medians and the median of non mature < the median of mature). The general procedure was: modify the data set to make the difference disappear, bootstrap the effect size and determine the p-value with a significance level = 0.05.

H0 : u1 <= u2
HA : u1 > u2

Viewership p-values results: Average Viewers (0.003<0.05), Peak Viewers (0.001<0.05), Views Gained (0<0.05) and Watch time Minutes (0.02<0.05).

Followers Activity p-values results: Followers (0.006<0.05), Followers Gained (0<0.05)

```
# non_mature["Followers"] <- non_mature["Followers"]
# - median_non_mature + median_mature - 1
# bs_dist <- tibble( d_in_medians=replicate( 1000,
#median( sample_nonmature() ) - median( sample_mature() ) ) )
# bs_dist[['extreme']] <- bs_dist[['d_in_medians']] > median_non_mature-median_mature
# ggplot( bs_dist, aes(y=d_in_medians,x="",col=extreme) ) +geom_quasirandom() +
# scale_color_manual( values = c( "TRUE" = "red", "FALSE" = "darkgray" ) )+
```

```
# geom_hline( yintercept = median_non_mature-median_mature ) +xlab("")+coord_flip()
```

Based on these results we can state that a channel having non-mature content gains:

- ~ 500 more Average viewers (95% CI: 169 to 859 ; p=0.003)
- ~ 5000 more Peak viewers (95% CI: 2504 to 8108 ; p= 0.001)
- ~ 1.800.000 more Views in a year (Gained)(95% CI: 1.232.977 to 2.626.370 ; p= 0)
- ~ 30.000.000 more Watch Time minutes (95% CI: -18.322.981 to 56.709.776 ; p=0.02)
- ~ 60.000 more Followers (95% CI: 18.773 to 102.559 ; p=0.006)
- ~ 40.000 more Followers in a year (Gained) (95% CI: 17.470 to 55.094 ; p=0)

2. Regression Analysis

Since the dependent variables of this analysis part are numeric in nature and the independent variable is categorical, we classify this task as creating a simple linear regression model (s) to measure the influence of the Mature attribute on the Audience Reactions variables.

```
data$Mature <- ifelse(data$Mature == "False", 0, 1)
model1 <- lm(Average.viewers ~ Mature, data = data)
model2 <- lm(Peak.viewers ~ Mature, data = data)
model3 <- lm(Views.gained ~ Mature, data = data)
model4 <- lm(Watch.time.Minutes. ~ Mature, data = data)
model5 <- lm(Followers ~ Mature, data = data)
model6 <- lm(Followers.gained ~ Mature, data = data)
#summary(model1)
```

D. Variable	R squared	Coefficient	CI 95%	p-value
Average.view	0.006	-1,639.3	-2882 -396	0.009
Peak.viewers	0.006	-11,853.9	-20,721 -2986	0.008
Views.gained	0.007	-5,212.472	-8,872,544 -1,552,399	0.005
Watch.Min	0.001	-56,389,985	~ - 137M ~ 24M	0.17
Followers	0.007	-169,121.2	-287,331 -50,911	0.05
Followers.g	0.008	-7,5081.19	-125,011 -25,150	0.03

Table 7. Simple Linear Regression

Based on the above results, the audience reactions for a channel with mature content is expected to be lower compared to a channel with non-mature content. However, we can observe a significant p-value for all metrics apart from Watch time minutes (0.17>0.05) and Followers

(0.05 = 0.05). Therefore, we do not have strong evidence to support here that the watched time minutes and the number of followers are lower for the channels generating mature content.

The uncertainty of the Watch time Minutes and Followers can also be confirmed by the wider confidence intervals of the coefficients in Table 7. The linear regression models also appear to have very low R squared scores, suggesting a weak prediction power and that only a small variance of the dependent variable(s) can be explained by the Mature attribute.

Apart from the Mature variable, other independent variables (Partnered, Language) were added to the analysis at this stage to measure their influence on the Audience Reactions variables, generating multiple linear regression models.

```
data$Mature <- ifelse(data$Mature == "False", 0, 1)
data$Partnered <- ifelse(data$Partnered == "False", 0, 1)
data$Language <- as.factor(data$Language)
model7 <- lm(Average.viewers ~ Mature + Partnered + Language, data = data)
model8 <- lm(Peak.viewers ~ Mature + Partnered + Language, data = data)
model9 <- lm(Views.gained ~ Mature + Partnered + Language, data = data)
model10 <- lm(Watch.time.Minutes.~ Mature + Partnered+ Language, data = data)
model11 <- lm(Followers ~ Mature + Partnered + Language, data = data)
model12 <- lm(Followers.gained ~ Mature + Partnered + Language, data = data)
```

Dependent Variable	Coef. MatureTrue	Coef. PartTrue	Coef. Lang English	p value	p value	p value	R squared
Average.view	-1,683	-107	-99.17	0.01	0.9547	0.9792	0.0006
Peak.viewers	-13,742	14,725	-12,51	0.003	0.2698	0.6419	0.01
Views.gained	-5,622,85	-2,555,71	1,286,622	0.0036	0.6461	0.3148	-0.002
Watch.Min	-74,788,572	147,806,523	256,583,613	0.0797	0.2293	0.3009	-0.004
Followers	-230,351	190,266	-42,728	0.0001	0.2772	0.9036	0.04
Followers.g	-79,833	43,286	-253454	0.0014	0.5472	0.0808	0.10

Table 8. Multiple Linear Regression

The multiple linear regression models build using Views Gained and Watch Time Minutes as dependent variables generated a negative adjusted R squared scores, suggesting that the chosen independent variables or predictors cannot provide any explanatory power over the performance metrics. Therefore, the interpretation of the associated coefficients was not performed in these cases as it can lead to misleading conclusions.

Based on the results in Table 8, the audience reactions (apart from Views Gained and Watch Time Minutes) numbers for a channel with mature content sees a significant decrease,

observation also made for the simple linear regression.

While the simple linear regression model's results on the Followers dependent variable determined no significance relationship between the mature content and the overall number of followers, the multiple linear regression suggested one. The number of followers of a channel with mature content is expected to be lower with 230,351. Due to the slightly higher R squared variable, the student decided to consider it as being significant. Moreover, the Partnered and Language (English) independent variables appear to have no significant effect on any performance metrics.

D. Variable	Coef. Language Chinese/p.	Coef. Language Japanese/p.	Coef. Language Korean /p.	Coef. Language Polish/p.	Coef. Language Russian/p.	Coef. Language Thai/p.
Followers. gained	-425117 0.006	-420455 0.01	-391696 0.008	-354926 0.03	-329271 0.02	-400329 0.02

Table 9. Multiple Linear Regression - Language

The outcome of the multiple linear regression model on Followers gained dependent variable also significantly suggests that the number of followers gained in a year of a channel with a Chinese/Japanese/Korean/Polish/Russian/Thai speaking language is expected to be less compared to a channel with Arabic speaking language. Moreover, the multiple linear regression models appear to still have very low R squared scores, suggesting a weak prediction power.

As mentioned in EDA, the quantitative variables have a right skewed distribution. According to Franco (2020), the presence of highly skewed variables (dependent or independent) can determine the non-normality of residuals, which is a violation of one of the linear regression assumptions. As not following the normality of the residuals can lead to inaccurate interpretations, we tested the assumption using the Shapiro-Wilk Test. The results in Table 10 indicate the non-normality of the residuals.

```
#residuals <- resid(model12)
#shapiro.test(residuals)
```

Model	Statistic	p-value
1	0.40031	<2.2e-16
2	0.51957	<2.2e-16
3	0.27086	<2.2e-16
4	0.50776	<2.2e-16
5	0.56786	<2.2e-16

Model	Statistic	p-value
6	0.53048	< 2.2e-16
7	0.43068	< 2.2e-16
8	0.56814	< 2.2e-16
9	0.27752	< 2.2e-16
10	0.53741	< 2.2e-16
11	0.61143	< 2.2e-16
12	0.59509	< 2.2e-16

Table 10. Shapiro-Wilk Test on Normality of Residuals

3. Overlapping Confidence Intervals Analysis

The Exploratory Data Analysis and Hypothesis Testing - Bootstrapping findings indicated that a channel having non-mature content has a higher viewership and followers activity, including the Watch time in minutes metric. However, the findings of the Regression Analysis indicated that a channel with mature content sees a decrease in viewership and followers activity metrics, apart from the Watch time in minutes variable.

Because of the lack of statistical evidence at this stage and also due to the wider confidence intervals received in EDA but also CDA for this evaluation variable, the student decided to also generate the confidence intervals between the two contents' medians to assess whether the values are indeed different for all audience reaction metrics.

To evaluate this matter, the student generated 95% confidence interval of the audience reactions medians using the Bootstrapping technique, more specifically, the interval between the 2.5/100 and 97.5/100 quantiles of the bootstrap distribution.

The confidence intervals overlapping observed in Table 11 may provide extra evidence for the lack of statistical significance between the medians of the two content types for Watch Time in minutes.

Dependent variable	95% CI - Mature	95% CI - Non - Mature	Overlap: Yes or No
Average.view	1.788 - 2318	2.358 - 2.790	No
Peak.viewers	11.118. - 15.203	16.789 - 20.549	No
Views.gained	4.569.532 - 5.569.918	6.547.592 - 7.610.128	No
Watch.Min	190.499.238 - 239.041.491	229.028.720 - 258.559.195	Yes
Followers	244.904 - 309.302	309.594 - 362.062	No
Followers.g	60.267 - 86.339	97.980 - 123.713	No

Table 11. Confidence Intervals 95%

The CDA findings do indicate some degree of statistical evidence to support the claim that a channel having non-mature content (matureFalse) has a higher viewership (including Watch Time). However due to the lack of statistical proof in Regression Analysis (matureTrue) and the overlapping between the medians of the two content types (matureTrue and matureFalse) for Watch Time, the claim on Watch Time was dropped.

After performing this action, the CDA findings now indicate that a channel with mature content has a lower viewership (apart from Watch Time) and lower followers activity metrics and that a channel with non-mature content has a higher viewership (apart from Watch Time) and higher followers activity.

II. Mature variable vs Partnered variable

1. Hypothesis Testing - Bootstrapping

The Exploratory Data Analysis results indicated that the mature content has a slightly slightly higher partnership rate than the non-mature content. However, the percentages indicated no real impact. As the ratio of proportions is close to 1 (97.83 % / 97.79 % = 1.0004), we checked the null hypothesis of ratio of proportions (H0 : u1 / u2 = 1) with a significance level = 0.05. A p-value of 0.467 was generated, indicating that we cannot reject the null hypothesis; we cannot reject that the two content types have the same partnership rate.

2. Regression Analysis

Since the dependent variable of this analysis part is categorical in nature and the independent variable is also categorical, we classify this task as creating a logistic regression model to measure the influence of the Mature variable on the probability of becoming a Twitch Partner.

```
data$Mature <- ifelse(data$Mature == "False", 0, 1)
data$Partnered <- ifelse(data$Partnered == "False", 0, 1)
glm_model <- glm(Partnered ~ Mature, #summary (glm_model)$coefficients
data = data, family = binomial(link = "logit"), control = list(maxit = 100))
```

Coefficient value	P value	AIC
0.01581	0.97	215.45

Table 12. Logistic Regression I

The outcome implies that a channel with mature content increases the likelihood of becoming Partnered comparing to channels with non-mature content. However the p-value indicates that this influence is not statistically significant ($0.97 > 0.05$).

Apart from the Mature variable, another independent variable (Language) was added at this stage to the analysis to measure its influence on Partnered variable, generating an additional logistic regression model.

To compare the predictive power of the first and second logistic regression models, the student compared them based on the Akaike Information Criterion values, since a lower score indicates a better performance. The second model has an AIC of 231 while the first one has a better one of 215.

The results of the second model implied that a channel with mature content decreases the likelihood of becoming Partnered but showed no statistical significance either. Language indicated no significance either over the odds of becoming Partnered. Overall, the CDA findings indicate that the influence of the Mature variable on Partnered is not significant, resulting into the same output suggested by EDA.

```
data$Partnered <- ifelse(data$Partnered == "False", 0, 1)
data$Mature <- ifelse(data$Mature == "False", 0, 1)
data$Language <- as.factor(data$Language)
glm_model2 <- glm(Partnered ~ Mature + Language, data = data,
family = binomial(link = "logit"), control = list(maxit = 100))
```

Coefficient value - Mature	P value	AIC
-0.1981	0.71	231.23

Table 13. Logistic Regression II

As Logistic Regression does not follow the assumptions of Linear Regression (2023), the test on the normality of the residuals was not performed at this stage.

III. Mature variable vs Streaming time (in minutes) variable

1. Hypothesis Testing - Bootstrapping

The Exploratory Data Analysis results indicated that the mature content determines more streaming time in minutes.

To check whether we are wrong about this statement, the null hypothesis of "no difference" was used (no difference in medians and the median of non mature > the median of mature), with a significance level = 0.05.

H0 : u1 >= u2

HA : u1 < u2

The received p-value of 0.45 indicates that we could not reject the null hypothesis and that we might be wrong about the EDA statement.

2. Regression Analysis

Since the dependent variable of this analysis part is numeric in nature and the independent variable is categorical, we classify this task as creating a simple linear regression model (s) to measure the influence of the Mature attribute on the Streaming time variable.

```
data$Mature<-ifelse(data$Mature=="False",0,1)
model13 <- lm(Stream.time.minutes. ~ Mature, data = data)
#summary(model13)
```

Coefficient value	p value	R squared
9,120	0.155	0.001

Table 14. Simple Linear Regression

Based on the above results, the Streaming Time in Minutes for a channel with Mature content is expected to be higher compared to a channel with non-mature content. However, p-value indicates no statistical significance. The linear regression model also appears to have very low R squared score (0.001), suggesting a weak prediction power and that only a small variance of the dependent variable(s) can be explained by the Mature attribute.

Apart from the Mature variable, other independent variables (Partnered, Language) were added to the analysis at this stage to measure their influence on the Streaming time variable, generating a multiple linear regression model.

```
data$Mature <- ifelse(data$Mature == "False", 0, 1)
data$Partnered <- ifelse(data$Partnered == "False", 0, 1)
data$Language <- as.factor(data$Language)
model14 <- lm(Stream.time.minutes.~ Mature + Partnered + Language,
data = data)
```

The results indicate that a channel with partnership decreases the streaming time but show no significant proof (see Table 15). Also, the multiple linear regression model indicates that the Streaming Time in Minutes for a channel with Mature content is expected to be higher but it's not statistically significant either.

Dependent Variable	Coef. value - Partnered	p-value	R squared
Streaming Time	-7,237	0.69	0.04

Table 15. Multiple LR. Partnered

However, the outcome also shows that the Streaming Time of a Twitch channel with a Chinese/French/Japanese/Portuguese/Thai speaking language is expected to be more compared to a channel with Arabic speaking language (see Table 16). A higher time increase appears to be for Portuguese and Thai languages. The multiple linear regression model has a better predictive performance (0.04) than the simple linear regression one (0.001).

Dependent Variable	Coef. Language Chinese/p.	Coef. Language Japanese/p.	Coef. Language French/p.	Coef. Language Portuguese/p.	Coef. Language Thai/p.
Streaming Time	99,025 0.01	93,134 0.04	80,673 0.03	107,102 0.005	114,169 0.01

Table 16. Multiple LR. Language

Overall, despite the pattern discovered in EDA, the CDA findings indicate that the influence of the Mature variable on Streaming Time in minutes is not significant.

Also, the Shapiro-Wilk Test results in Table 17 indicate the non-normality of the residuals.

```
# residuals <- resid(model14)
# shapiro.test(residuals)
```

Model	Statistic	p-value
13	0.75179	< 2.2e-16
14	0.77214	< 2.2e-16

Table 17. Shapiro-Wilk Test on Normality of Residuals

5. Recommendation

The 3 questions of this case study were approached using Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA) statistical methods. The techniques implied

first inspecting, exploring the available data, forming hypotheses and then using different tools (Hypothesis Testing Bootstrapping and Regression Analysis - Simple/Multiple Linear Regression and Logistic Regression) to prove that these hypotheses can be true.

However, it is important to emphasize that the recommendations (5.1 Guidance) may not reflect the reality and that they were mainly generated based on the received statistical directions of the data. The limitations of the findings will be discussed in the "Limitations" section, followed by the "Next steps" unit, to address them.

5.1 Guidance

Based on the results obtained from EDA and CDA stages and the given inquiries, the following recommendations can be given:

I. Should I focus on mature content for a grown-up audience? What are the consequences if I decide to do so in terms of audience reactions? Is the effect constant across all different ways one can think about "performance".

No, you should focus on non-mature content instead. The CDA findings (Hypothesis Testing - Bootstrapping) indicate some degree of statistical evidence to support the claim that a channel having non-mature content has a higher viewership (apart from Watch Time) and higher followers activity. Also, the CDA results (Regression Analysis) indicate that a channel with mature content has a lower viewership (apart from Watch Time) and lower followers activity. As a result, if the client decides to focus on mature content, he might receive a lower viewership and followers activity, a performance decrease which would have a big negative impact on having a successful channel. Concerning the effect across all different performance metrics, it varies in strength for all variables, matter which will be further addressed in the "Next Steps" section. The performance metrics may imply the following:

Average Viewers: the non-mature content has a better ability to attract the audience in a particular stream session

Views Gained: a channel generating non-mature content collects a considerably higher number of overall views (e.g. from live, archive streams) during a year

Peak Viewers: there are times when channels with non-mature content do manage to attract a larger number of viewers in a stream

Followers Gained / Followers: a channel with non-mature content may generate an increase in followers, who can be represented by the young users who tend to be more active on such platforms possibly due to their more free time and who watch this type of content as it aligns more with their entertainment preference

Watch Time in minutes: it was not statistically proven that the choice of focusing on mature or non-mature content has an impact on Watch Time in minutes. Consequently, we cannot

say that a non-mature channel has a better ability to determine a longer watch time for the overall channel

An extra recommendation at this stage would be: if the client decides to use Arabic as a speaking language to not switch to either Chinese/Japanese/Korean/Polish/Russian/Thai as he might see a decrease in Followers gained in a year.

II. Does this choice lower or increase my chance of becoming a Twitch partner?

It slightly increases the chance according to EDA but it is not statistically proven in CDA. The EDA stage shows that a channel with mature content has a slightly higher partnership rate, meaning an increased chance of becoming a Twitch partner compared to non-mature. However, the CDA findings indicated no statistical significance. We cannot reject that the content types have the same partnership rate and Regression Analysis indicated no significant influence. As a result, the content nature - mature could not be proven to have any impact on the chance of becoming a Twitch Partner.

III. As you know, I am sluggish: Is the effect of the minutes I stream larger or smaller in mature content?

It is larger according to EDA but it is not statistically proven in CDA. We cannot reject that a channel with non-mature content has a higher or equal Streaming Time than the mature content. Also, Regression Analysis indicated no significant influence. As a result, the content nature could not be proven to have any impact on increasing or decreasing the Streaming time in minutes.

Overall recommendation from the available data: Your channel should focus on generating non-mature and not mature content to gain higher viewership and following activity; this statement represents a statistical direction of the available data and it is not conclusive. Focusing on mature content might increase the chance of becoming a Twitch Partner and increase the Streaming Time; however these two statements remain assumptions that need to be further analyzed to assess their accuracy.

5.2 Limitations

- Class Imbalance

The student considers that the imbalanced distributions of Partnered variable (97% - TRUE) and Mature variable (77% - FALSE) may pose certain risks to the analysis. Their imbalance may either suggest the reality of a main trend about the majority class or indicate a bias, that is in reality influenced by external factors, so the majority class may not be the only factor contributing to success. Regarding the minority class, the imbalance problem may pose limitations in capturing its unique characteristics.

Concerning the answer of the first question, the findings might be influenced, be biased towards the majority class of Mature variable: FALSE, possibly due to the policies of the Twitch

Platform (only allowing mature content in specific conditions) or the preferance of users. The study may also not capture the unique characteristics of the mature content as it does not have sufficient representation and offer uncertain, unrepresentative, biased findings (Risk: Inaccurate results).

Regarding the second question, the imbalance issue might imply that no matter what content the client picks, there will be the same chance of getting a Partnership:TRUE. The bias towards the majority class might be due to the fact that streamers with a long registration on Twitch get partnerships (only not approving partnership for almost new channels) (Risk: Misinterpretation of the influence of channel content on partnership). The minority class, Partnership:FALSE also does not have sufficient representation to offer accurate results.

Ultimately, regarding the third question, even though the discovered pattern might not be biased towards the majority class, Mature variable: FALSE, the imbalance issue still can pose problems, as the limited representation of the minority, Mature variable: TRUE can lead to misleading findings (Risk: Inaccurate results).

- Wide Confidence Intervals

Throughout this case study, 95% confidence intervals were used to assess the magnitude of impact of independent variables in the growth or decrease of dependent variables. While some of them were narrow, indicating a precise population estimation, quite a few were wide, suggesting the uncertainty in interpretation.

For example, when analyzing the relationship between Mature and Audience Reactions, the number of Followers for non-mature content is uncertain in growth amount comparing to the other variables; the number of Followers for mature content is uncertain in decreased amount comparing to the other variables. The Watched Time for mature/non-mature content is extra uncertain in growth or decrease comparing to non-mature/mature content, implying uncertainty on its true direction and interpretation.

As a result, the wide confidence intervals need to be interpreted with carefulness and awareness to not make inaccurate decisions.

- Time Misinterpretation

This case study was conducted on the Top 1000 Twitch streamers in 2020. As the current year of this analysis is 2023, the student considers that the findings generated from this data might not show the reality of the best performing Twitch channels in 2023. Over the span of 3 years, there could have been various modifications to the nature and quantity of success indicators for a Twitch channel. For instance, in 2020, the success of a streaming channel could have been only based on a number of overall Followers and a certain amount of Average Views. However, in 2023, the success could be defined based on a lower amount of Followers and Average Views but also based on Streaming Time. Thus, the rankings of the top channels could have easily changed as well according to the updated level of success.

As a result, the findings of this analysis might not be accurate to achieve the client's goal as they would give more insights and recommendations into what he should have focused on to gain success in 2020, rather than 2023.

- Non-normality of the residuals

The normality of the residuals was assessed via the Shapiro-Wilk Test for the simple and multiple linear regression models of the first and third inquiries. Since the results of the test suggested a violation of linear regression's assumptions, the findings and interpretations may not be valid, namely: that a channel with mature content has a lower viewership and followers activity and that focusing on mature content increases the Streaming Time.

5.3 Next Steps

5.3.1 Confidence Intervals Interpretation

The magnitude of impact of independent variables in the growth/decrease of dependent variables (95% CI) needs to be further discussed with the client or domain experts to:

- check whether the values align with the project's goal (e.g. check whether the constant effect increase/decrease of the dependent variable is enough when focusing on non-mature/mature content)
- examine external factors (e.g. marketing strategies, trends) of the Twitch platform to understand the context of the assignment better

5.3.2 Further Analysis on Assumption

According to the student, the weak EDA hypothesis of Question 2 - the Mature content having a higher partnership rate than Non-mature does not align with a potential accurate pattern, considering the recommendation of Question 1, external sources but also the overall imbalance problem of the variables.

Regarding the imbalance situation of not having a main trend for Partnered:TRUE, the student considers that a more reliable pattern would have needed to be indicated by a considerably higher Partnership rate, lower Non-Partnership rate for the Mature content and statistical significance in Hypothesis Testing - Bootstrapping and Regression Analysis.

According to an up-to-date external source (2023), there are 2 possible ways to gain channel profitability and other features: by being an Affiliate or a Partner. The main difference between these 2 partnerships is that the application requirements imply a difference in viewership, stream time and followers activity obtained in a month.

As a result, we can form an assumption based on the potential accuracy of the answer of question 1, that can be further analyzed in the future by collecting up-to-date data: a channel generating Non-mature content has a higher chance of becoming a Twitch Partner since its implied higher audience reaction metrics suggest a higher chance for a partnership approval.

However, it is worth critically thinking that despite the imbalance situation of having a main trend or not, there might be possible undiscovered external factors that could influence the partnership decision, apart from content nature and audience reactions (e.g. professionalism and content quality), that would need to be researched first and collected, to prove any statement.

5.3.3 Non-normality of Residuals - Variables Transformation

According to Franco (2020), if the dependent or independent variables of an analysis are highly skewed, we should normalize them before building linear regression models, in order to attempt to determine the normality of the residuals. Since this assumption is violated in this analysis, possible transformation techniques that could be applied on the quantitative variables to tackle skewness and thus improve the case study results, would be Log Transformation or Boxcox Transformation (Saxena, 2020).

5.3.4 New Data Collection

New data of the Top Twitch streamers in 2023 needs to be collected to receive more accurate findings about what decisions to make in order to have a successful Twitch channel. Such ranking information can be found on TwitchTracker website, a platform offering live statistics of streamers, including viewership, streaming time and followers activity metrics. This data can be collected via an API, using R/Python programming languages. In this manner, the time misinterpretation limitation can be addressed.

Concerning the sample amount, the student considers that there are 2 methods that can be used to address the other 2 limitations of this analysis (class imbalance and wide confidence intervals).

The first one is to add more channels from the ranking, until the distribution of the imbalanced variables becomes more balanced. In this way, we could analyze the success factors over a larger number of channels. However, if we increase it and add less and less performing players, we will have a broader view, possibly leading to the use of patterns that will generate less success impact for the client.

The second approach could be applied after using the first method or not, depending on the preference of the client. It implies comparing success patterns between 2 samples: top channels and the rest of the observations (average performing channels), to confirm that the characteristics of best players are indeed unique.

The Language and Mature variables could be added to the new data set by applying machine learning techniques on the past streams of each channel.

An additional change to the data would be to only include the channels with English speaking language as this data might lead to more accurate insights that can better fulfill the client's goals.

Having the main audience reaction, streaming metrics, language, and mature type, new variables can be added from various sources, to improve the analysis:

I. Mature variable vs Audience Reaction

New independent variables that can be used to analyze their influence together with Mature content in predicting the dependent variable, can be collected via the profile page of a Twitch Streamer.

Via an API, extra data such as bio description, email, streaming schedule habits, the categories and the number of past streams can be gathered.

Regarding the bio description, it would be interesting to analyze the relationship between the description updates and the changes in viewership and follower activity metrics. In this manner, Natural Language Processing techniques can be used to discover any keywords or phrases that might capture the audience's attention and lead to an increase in reactions. The categories of past streams can also be assessed to check whether there is a relationship between the choice of a category and audience reactions. Moreover, having the number of past streams, we could determine the Average Watch time from the Total Watch time to check which content (Mature/Non-mature) has a better ability to engage the audience for a longer period once they enter a stream.

II. Mature variable vs Partnered variable

The external factors that could influence the partnership decision, apart from content nature and audience reactions could be the signs of content quality, professionalism and social media activity.

The content quality and professionalism metrics can be collected from the audience's feedback via online surveys, interviews or focus groups. In this manner, we can assess the relationship between the strengths and weaknesses of a streamer with the partnership rate. Regarding the social media activity, a large amount of data can be collected via API platforms to analyze the audience's view on the streamer's competence and content: overall sentiment (positive/negative) (obtained by analyzing comments, tweets or conversations) and engagement rate.

6. Conclusion

In conclusion, this paper presents a data analytics case study of the Top 1000 Twitch streamers, supporting the decision-making process of an individual who is about to enter the online content production field. The 3 questions of the assignment were approached using Exploratory Data Analysis and Confirmatory Data Analysis statistical techniques, providing guidance, limitations and possible next steps.

7. References

Assumptions of logistic regression. Statistics Solutions. (2021, August 11). https://

www.statisticssolutions.com

/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

Saxena, A. (2020, October 26). *How transformation can remove skewness and increase accuracy of linear regression model*. Medium. https://anshikaaxena.medium.com/how-skewed-data-can-skrew-your-linear-regression-model-accuracy-and-transfromation-can-help-62c6d3fe4c53

Franco, G. (2020, March 3). *Linear regression: Should dependent and independent variables be distributed normally?*. StatsImprove. https://www.statsimprove.com/en/linear-regression-should-dependent-and-independent-variables-be-distributed-normally/?fbclid=

IwAR2bUxco5OcUgDoiJ3CvMi9P7MY0TYjQ0F84HgEauI_51J8uBXopHVU1PIU

Prof, C. (2022, August 5). *5 ways to check the normality of residuals in R [examples]*. CodingProf.com. https://www.codingprof.com/5-ways-to-check-the-normality-of-residuals-in-r-examples/?utm_content=cmp-true&fbclid=IwAR3qcNiWSKtjLaGw3M-_5GL-ZnI0wfgk8r9I1yy2qCkq2Ic9a-AWlzfWhNM

Radboud University. Data Analysis Course. Slides

Twitch affiliate vs partner - A detailed rundown of differences. TrueList. (2023, January 7). https://truelist.co/blog/twitch-affiliate-vs-partner/

8. Appendix

```
num_vars <- sum(sapply(data, is.numeric))
rows <- ceiling(sqrt(num_vars))
cols <- ceiling(num_vars / rows)
par(mfrow = c(rows, cols))
par(mar = c(2, 2, 1, 1))
for (col in names(data)) {
  if (col != "Mature" && col != "Partnered" && is.numeric(data[[col]])) {
    hist(data[[col]], main = col, xlab = "Value", col = "skyblue")
  }
}
```

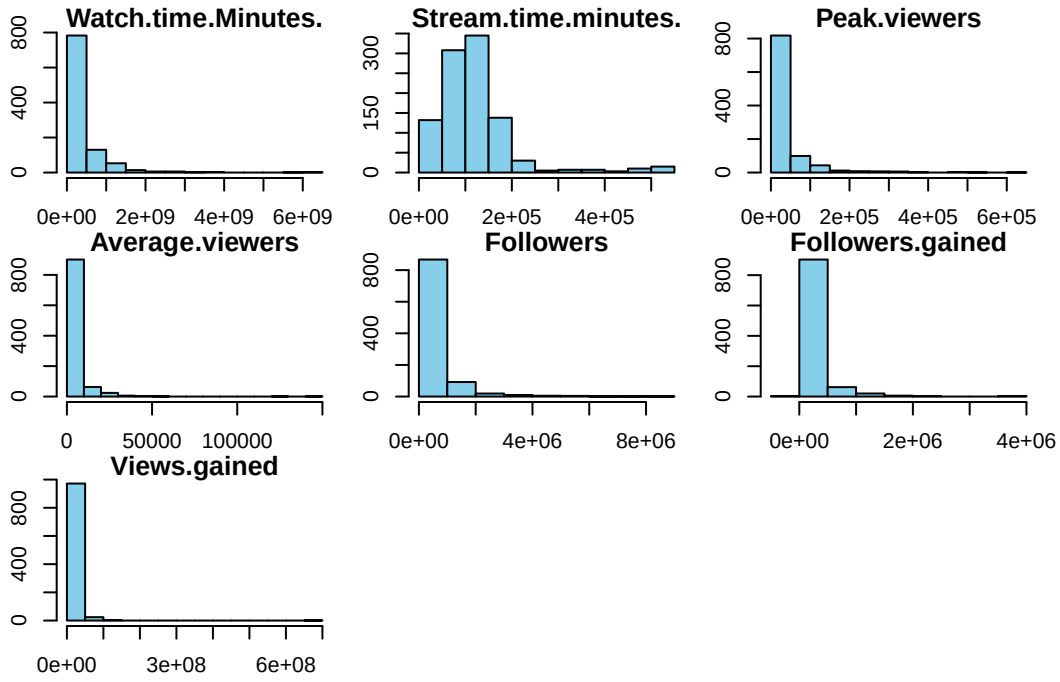

Figure 1. Univariate distribution each numerical column

```
data <- read.csv("twitchdata-update.csv")
ggplot(data, aes(x = Mature, y = Average.viewers, fill = Mature)) +
geom_bar(stat = 'summary', fun = 'median') +
stat_summary(geom = 'text', fun = 'median', aes(label = after_stat(y),
group = Mature), vjust = -0.1)
```

Figure 2. Average viewers vs Mature

```
data <- read.csv("twitchdata-update.csv")
ggplot(data, aes(x = Mature, y = Views.gained, fill = Mature)) +
geom_bar(stat = 'summary', fun = 'median') +
stat_summary(geom = 'text', fun = 'median', aes(label = after_stat(y),
group = Mature), vjust = -0.1)
```

Figure 3. Views Gained vs Mature

```
data <- read.csv("twitchdata-update.csv")
ggplot(data, aes(x = Mature, y = Peak.viewers, fill = Mature)) +
geom_bar(stat = 'summary', fun = 'median') +
stat_summary(geom = 'text', fun = 'median', aes(label = after_stat(y),
group = Mature), vjust = -0.1)
```

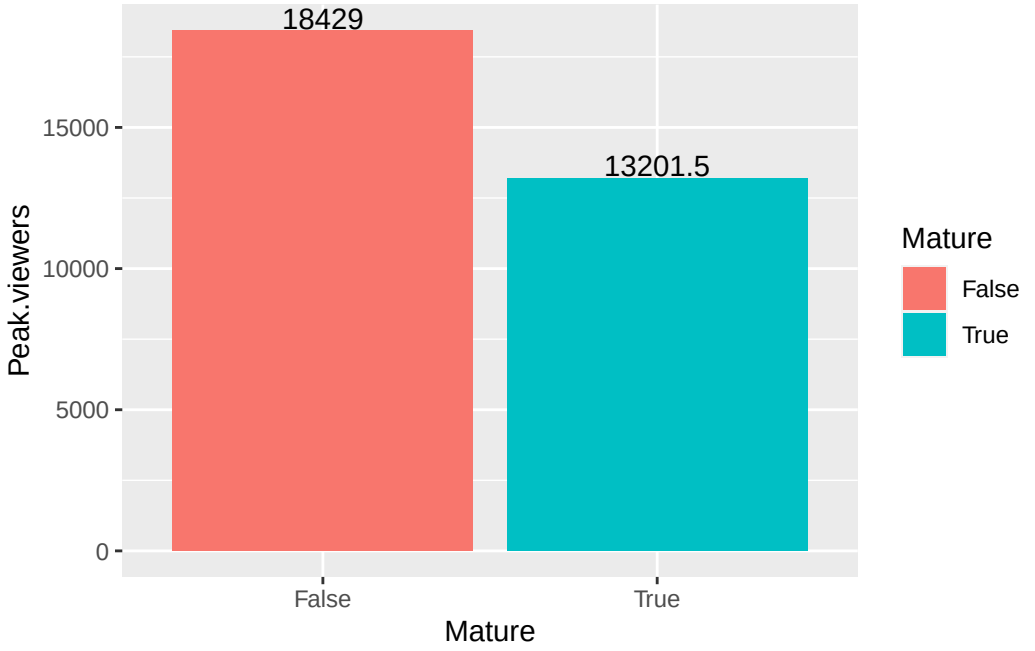

Figure 4. Peak Viewers vs Mature

```
data <- read.csv("twitchdata-update.csv")
ggplot(data, aes(x = Mature, y = Watch.time.Minutes., fill = Mature)) +
geom_bar(stat = 'summary', fun = 'median') +
stat_summary(geom = 'text', fun = 'median', aes(label = after_stat(y),
group = Mature), vjust = -0.1)
```

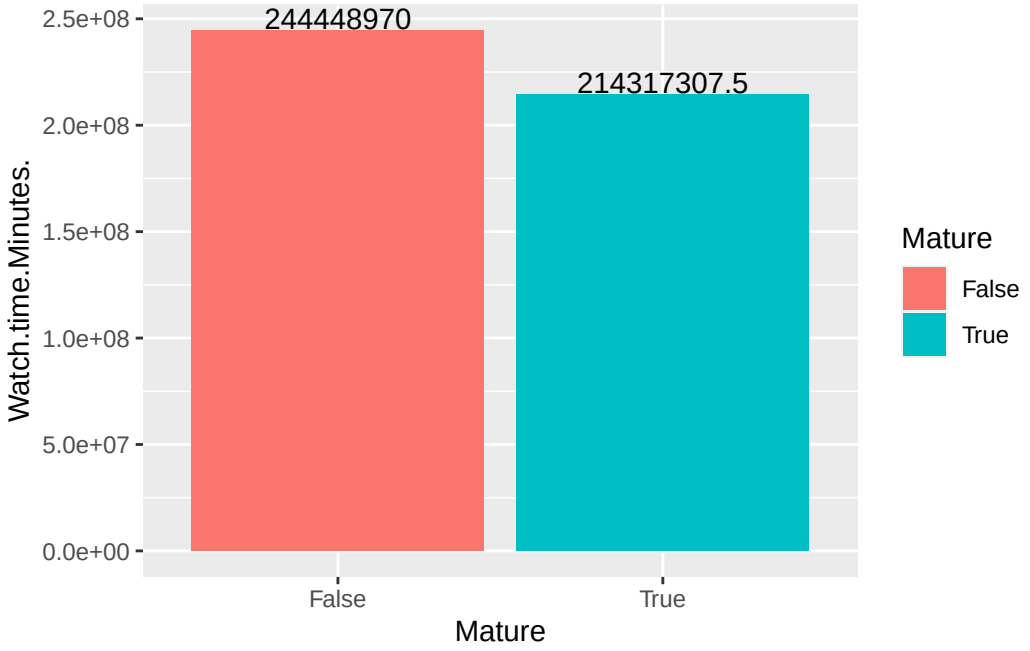

Figure 5. Watch time Minutes vs Mature

```
data <- read.csv("twitchdata-update.csv")
ggplot(data, aes(x = Mature, y = Followers, fill = Mature)) +
geom_bar(stat = 'summary', fun = 'median') +
stat_summary(geom = 'text', fun = 'median', aes(label = after_stat(y),
group = Mature), vjust = -0.1)
```

Figure 6. Followers vs Mature

```
data <- read.csv("twitchdata-update.csv")
ggplot(data, aes(x = Mature, y = Followers.gained, fill = Mature)) +
geom_bar(stat = 'summary', fun = 'median') +
stat_summary(geom = 'text', fun = 'median', aes(label = after_stat(y),
group = Mature), vjust = -0.1)
```

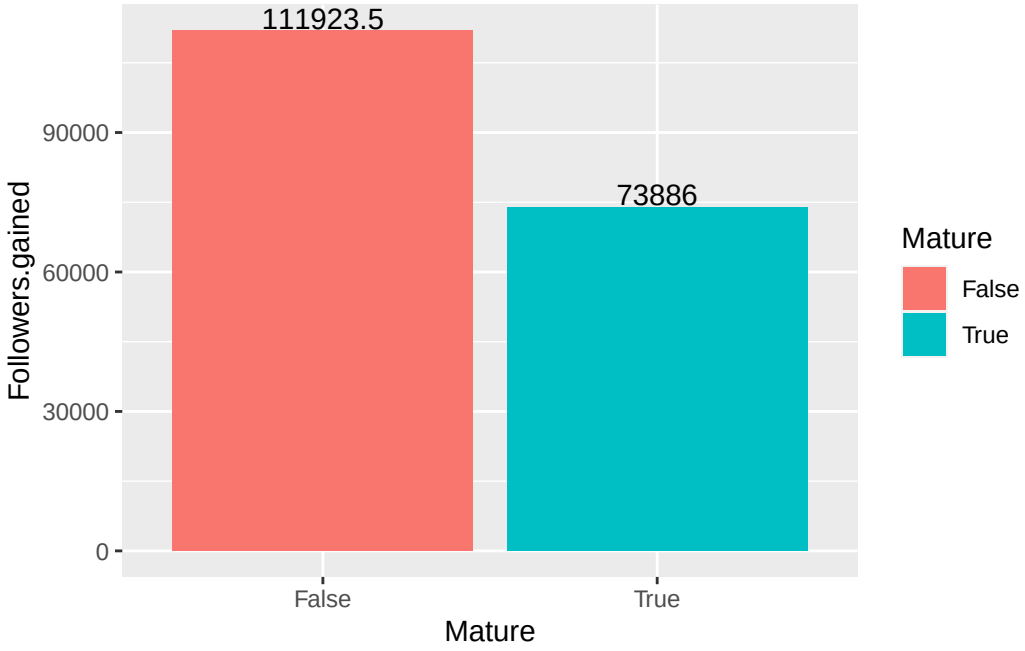

Figure 7. Followers Gained vs Mature