SOCIAL MEDIA HABITS:

Can we predict social media use?

Derrick Chun, Andrew Gibson, Linh Lai, Jiakai Lin, Emily Zhao

MA 214

Thomas Enkosky

November 26, 2024

PART 1: 2
PROBLEM 2
BACKGROUND 3
VARIABLES 3
CONTRIBUTIONS 3
PART 2: 3
DATA COLLECTION METHOD 3
SAMPLE DATA 4
CONFIDENCE INTERVALS 4
ANALYSIS 6
CONTRIBUTIONS 6
PART 3: 6
INTRODUCTION 6
SCATTER PLOTS 7
ANALYSIS 9
RESIDUAL PLOTS 9
ANALYSIS 10
PROBLEMS 10
CONTRIBUTIONS 10
PART 4: 10
MULTICOLLINEARITY 10
ANALYSIS 11
MODEL FIT 11
ANALYSIS 12
VARIABLE SELECTION 12
STEPWISE REGRESSION 13
ANALYSIS 13
GENERALIZED LINEAR MODEL 13
ANALYSIS 14
CONFIDENCE AND PREDICTION INTERVALS 14
ANALYSIS 15
SUMMARY 16
CONTRIBUTIONS 16
SOURCES 16

PART 1:

PROBLEM

This statistics project is designed to analyze social media use. Our research question is as followed:

- Can we predict time spent on social media based on additional information about the user?

BACKGROUND

In today's age, nearly anyone is capable of accessing social media. However, not everyone spends the same amount of time on these platforms. Some people barely use it at all, while others spend most of their day on social media. This disparity may seem random at first, but we believe there may be external factors that influence someone's time spent on social media. Our initial hypothesis was that social media usage would be higher among young, female users who are students and come from a higher income background. This project will construct a multiple regression model, which will predict daily hours on social media based on the inputted variables, which consist of quantitative and qualitative (categorical) data.

VARIABLES

To effectively analyze this dataset, we had to define the variables, listed below. We decided to remove variables that were either confusing or not relevant to the question. We trimmed down the dataset to include 12 variables. We categorized them below:

- Dependent/Response: Total time spent on social media
- Independent/Predictor:
    - Quantitative: age, income, number of sessions, engagement, time spent on video, number of videos watched, scroll rate
    - Qualitative: gender, profession, platform, watch reason

CONTRIBUTIONS

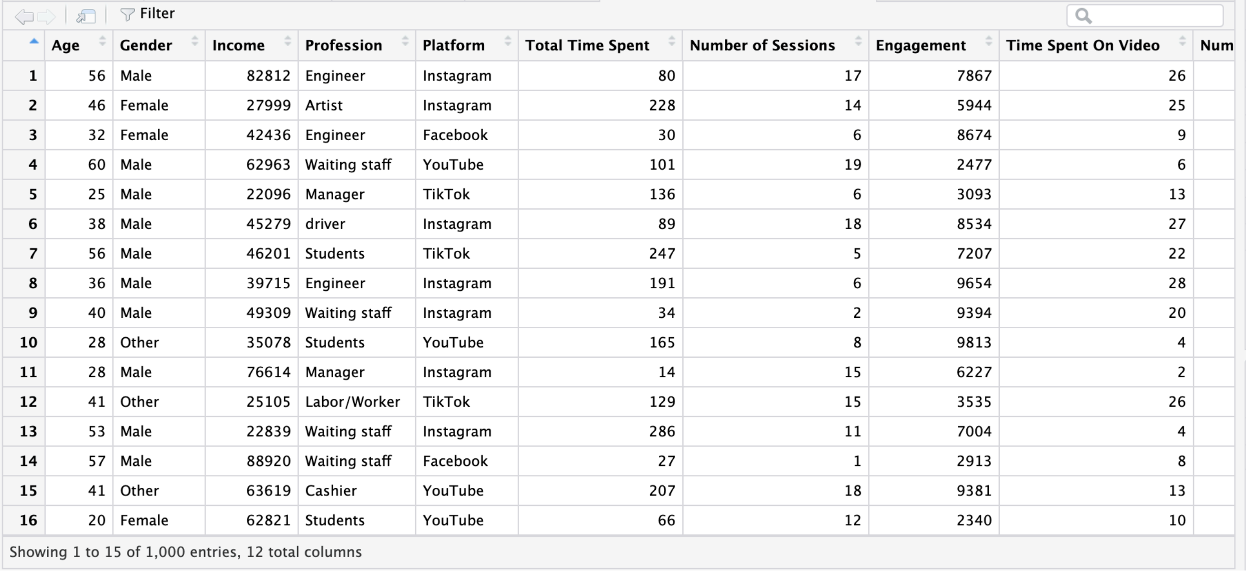
Everyone worked together to determine the project goal. Emily Zhao worked on determining the predictor variables and brainstorming ideas for research questions. Andrew Gibson brainstormed research questions and variables to analyze, and wrote the hypothesis. Derrick Chun researched variables related with social media use, and added quantitative and qualitative variables, and edited the hypothesis. Linh Lai brainstormed project ideas. Jiakai Lin found the dataset and brainstormed project ideas. All members will contribute group milestones by doing a fair share of coding in R.

PART 2:

DATA COLLECTION METHOD

Our team selected a dataset off the data aggregation website Kaggle, named "Social Media Menace." This dataset contains 31 variables across 1000 observations. It was generated using Python libraries NumPy and Pandas to mimic real world social media usage.

SAMPLE DATA



	Age	Gender	Income	Profession	Platform	Total Time Spent	Number of Sessions	Engagement	Time Spent On Video	Num
1	56	Male	82812	Engineer	Instagram	80	17	7867	26	
2	46	Female	27999	Artist	Instagram	228	14	5944	25	
3	32	Female	42436	Engineer	Facebook	30	6	8674	9	
4	60	Male	62963	Waiting staff	YouTube	101	19	2477	6	
5	25	Male	22096	Manager	TikTok	136	6	3093	13	
6	38	Male	45279	driver	Instagram	89	18	8534	27	
7	56	Male	46201	Students	TikTok	247	5	7207	22	
8	36	Male	39715	Engineer	Instagram	191	6	9654	28	
9	40	Male	49309	Waiting staff	Instagram	34	2	9394	20	
10	28	Other	35078	Students	YouTube	165	8	9813	4	
11	28	Male	76614	Manager	Instagram	14	15	6227	2	
12	41	Other	25105	Labor/Worker	TikTok	129	15	3535	26	
13	53	Male	22839	Waiting staff	Instagram	286	11	7004	4	
14	57	Male	88920	Waiting staff	Facebook	27	1	2913	8	
15	41	Other	63619	Cashier	YouTube	207	18	9381	13	
16	20	Female	62821	Students	YouTube	66	12	2340	10	

Showing 1 to 15 of 1,000 entries, 12 total columns

CONFIDENCE INTERVALS

Quantitative Data:

1. Total Time Spent (response variable): (146.1963, 156.6157)
data: Predicting_Social_Media_Usage$`Total Time Spent`
t = 57.031, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: (146.1963 156.6157)
sample estimates: mean of x = 151.406

2. Age: (40.1484 41.8236)
data: Predicting_Social_Media_Usage$Age
t = 96.022, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: (40.1484 41.8236)
sample estimates: mean of x = 40.986

3. Income: (58051.27, 60997.16)
data: Predicting_Social_Media_Usage$Income
t = 79.302, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: (58051.27 60997.16)
sample estimates: mean of x = 59524.21

4. Number of Sessions: (9.679126, 10.346874)
data: Predicting_Social_Media_Usage$`Number of Sessions`
t = 58.851, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0

95 percent confidence interval: (9.679126 10.346874)
sample estimates: mean of x = 10.013

5. Engagement:
data: Predicting_Social_Media_Usage$Engagement
t = 54.303, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: (4816.577 5177.741)
sample estimates: mean of x = 4997.159

6. Time Spent on Video:
data: Predicting_Social_Media_Usage$`Time Spent On Video`
t = 57.742, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: (14.46415 15.48185)
sample estimates: mean of x = 14.973

7. Number of Videos Watched: (24.37742 26.11858)
data: Predicting_Social_Media_Usage$`Number of Videos Watched`
t = 56.911, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: (24.37742 26.11858)
sample estimates: mean of x = 25.248

8. Scroll Rate: (47.96214 51.58586)
data: Predicting_Social_Media_Usage$`Scroll Rate`
t = 53.908, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: (47.96214 51.58586)
sample estimates: mean of x = 49.774

Qualitative Data:

1. Gender:
Chi-squared test for given probabilities:
data: gender_counts
X-squared = 184.33, df = 2, p-value < 2.2e-16

2. Profession:
Chi-squared test for given probabilities:
data: profession_counts
X-squared = 435.72, df = 8, p-value < 2.2e-16

3. Platform:
Chi-squared test for given probabilities:
data: platform_counts
X-squared = 5.624, df = 3, p-value = 0.1314

4. Watch Reason:
Chi-squared test for given probabilities:
data: watch_reason_count

```
X-squared = 102.98, df = 3, p-value < 2.2e-16
```

ANALYSIS

An existing article analyzing social media usage and comparing them to characteristics of people and their hobbies reinforces some of the data set. The study has a correlation matrix, and also tables that display social media engagement and usage frequency compared with auxiliary variables, such as gender, education level, and income. Due to the scope of our dataset, we are only looking at the frequency data, not the engagement data, as it does not count videos watched or scroll rate as engagement. As a result, we were not able to find more data on our variables for scroll rate, number of videos watched, or time spent per video. We also did not find any other data for watch reason or platform.

The tables are split into minimal, moderate, and high frequency. For gender, women were identified as 1, and men were identified as 0. The average for those with high frequency of usage was 0.52, and 0.34 for low frequency, which reinforces our Chi-Squared test that there is possible interdependence between gender and social media usage. Income was categorized from 2 to 8, and low frequency was 3.73, while high frequency was 3.48, with moderate frequency being in the middle. The study does not have categories based on job profession, but rather education level, which can somewhat be a similar predictor. Those who attended “some college” or had a bachelor’s degree had higher usage rates compared to those who only completed high school or had an associate’s degree. The study’s mean participant age was 23.06 with 1.91 standard deviation, so we have an average age that is older than their dataset. We believe that our data set is more representative of the population because of this. Additionally, their survey only gathered 249 participants, so our data set should be more representative due to having an even larger sample size.

CONTRIBUTIONS

Emily Zhao worked on data collection methods. Andrew Gibson helped select variables to use, formatted, edited and conducted confidence interval tests. Derrick Chun selected what variables to use for our project, conducted confidence interval and chi-square tests and created the updated CSV file that everyone will use. RStudio was used to compute the confidence intervals for all the quantitative variables and conducted the Chi test for the qualitative variables. Linh Lai wrote the introduction. Jiakai Lin helped select variables to use and conducted confidence interval tests.

PART 3:

INTRODUCTION

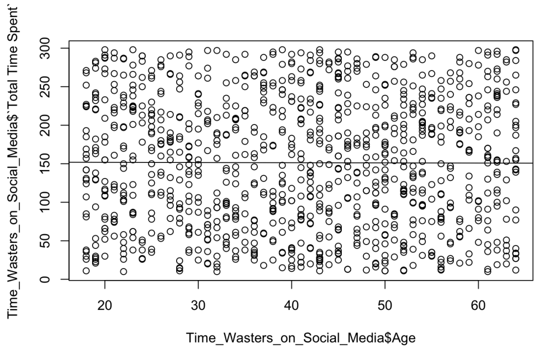
Our study aims to analyze social media patterns by exploring how various user characteristics might predict the total time spent on social media. Social media is a central part of many people’s everyday lives, but the amount of time spent on it varies from person to person. Our project investigates whether factors such as age, income, gender, and profession, along with behaviors like the number of sessions and videos watched, can predict this social media usage time. We believe these factors are relevant to social media usage based on existing research and logical connections to user behavior. For example, age is a key factor, as younger individuals are generally more immersed in social media, using it for entertainment, communication, and even professional networking. Gender is also relevant as studies suggest that women tend to use social media more frequently for social interaction and content consumption. From looking at the data, we hypothesize that younger individuals, females, students, and those with higher incomes are more likely to spend more time on social media.

SCATTER PLOTS

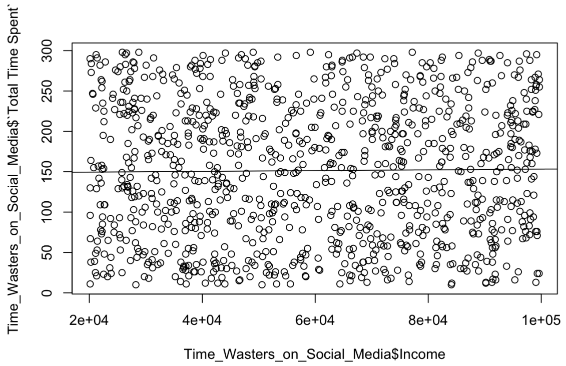
In all of the scatterplots, Y is the time spent on social media.

Scatterplots:

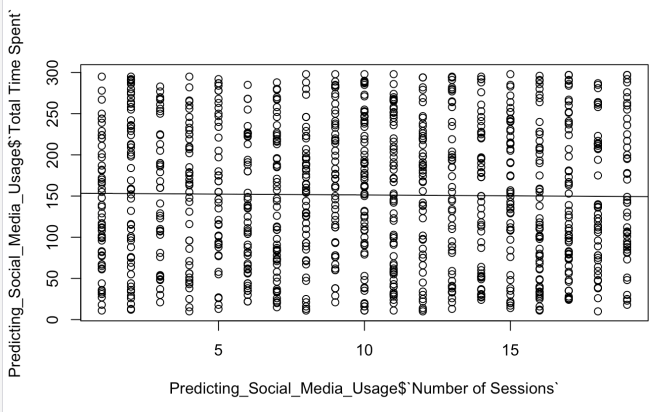
1. Time spent on social media vs age (x1)
$y = 152.468 - 0.02591x1$



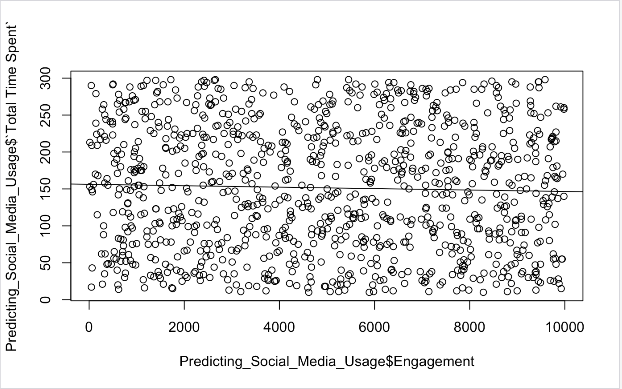
2. Time spent on social media vs income (x2)
$y = 148.6 + 0.00004713x2$



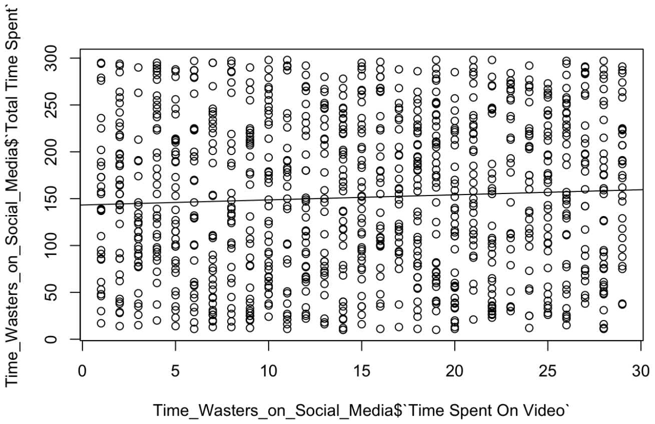
3. Time spent on social media vs number of sessions (x3)
$y= 153.5739 - 0.2165x3$



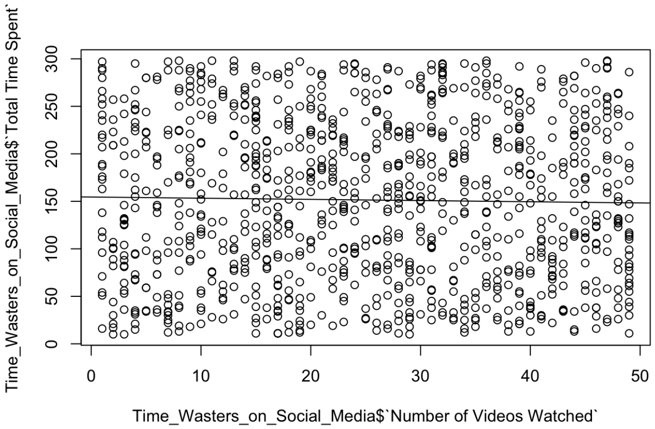
4. Time spent on social media vs engagement(x4)
$y= 156.3 - .0009735x4$



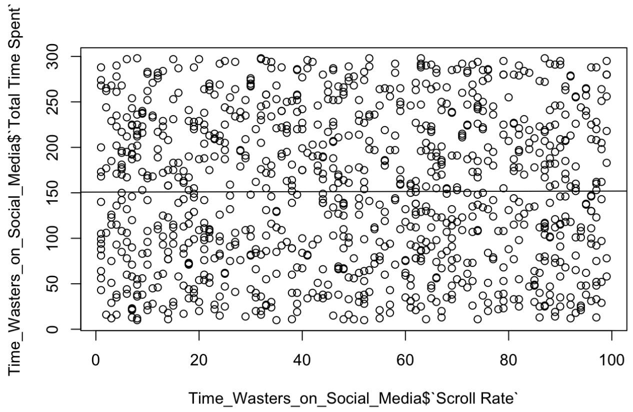
5. Time spent on social media vs time spent on video (x5)
y = 143.2326 + .5459x5



6. Time spent on social media vs number of videos watched (x6)
y = 154.57 - .1255x6



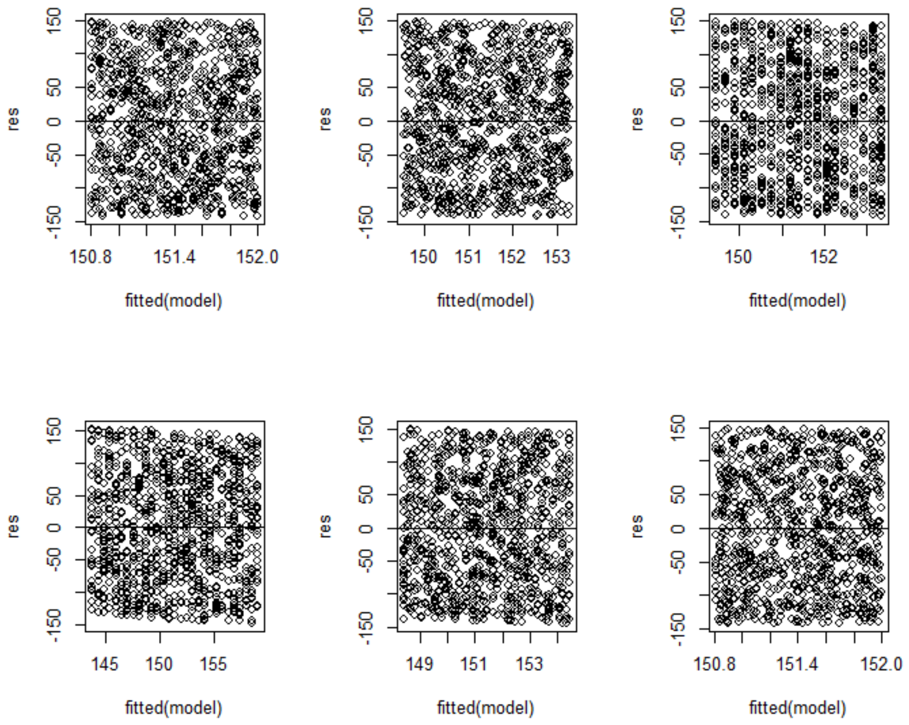
7. Time spent on social media vs scroll rate (x7)
y = 150.803 + 0.01211x7



ANALYSIS

Overall, all of our variables seem to be randomly scattered and uncorrelated to time spent on social media. However, there is the chance for multivariable regression to show some correlation, so we will continue investigating.

RESIDUAL PLOTS



ANALYSIS

The assumptions of the linear regression model are satisfied. There is a random pattern in the residuals suggesting the linearity assumption is likely met. The spread of residuals is roughly the same across the range of fitted values which supports the homoscedasticity assumption. There is no clear pattern or clustering which suggests the independence assumption is satisfied and the residuals appear fairly symmetric which indicates the satisfies the normality assumption. We did not transform or remove any of the variables, as we did not see any correlation in the initial scatter plots. It was not worth it to do so.

PROBLEMS

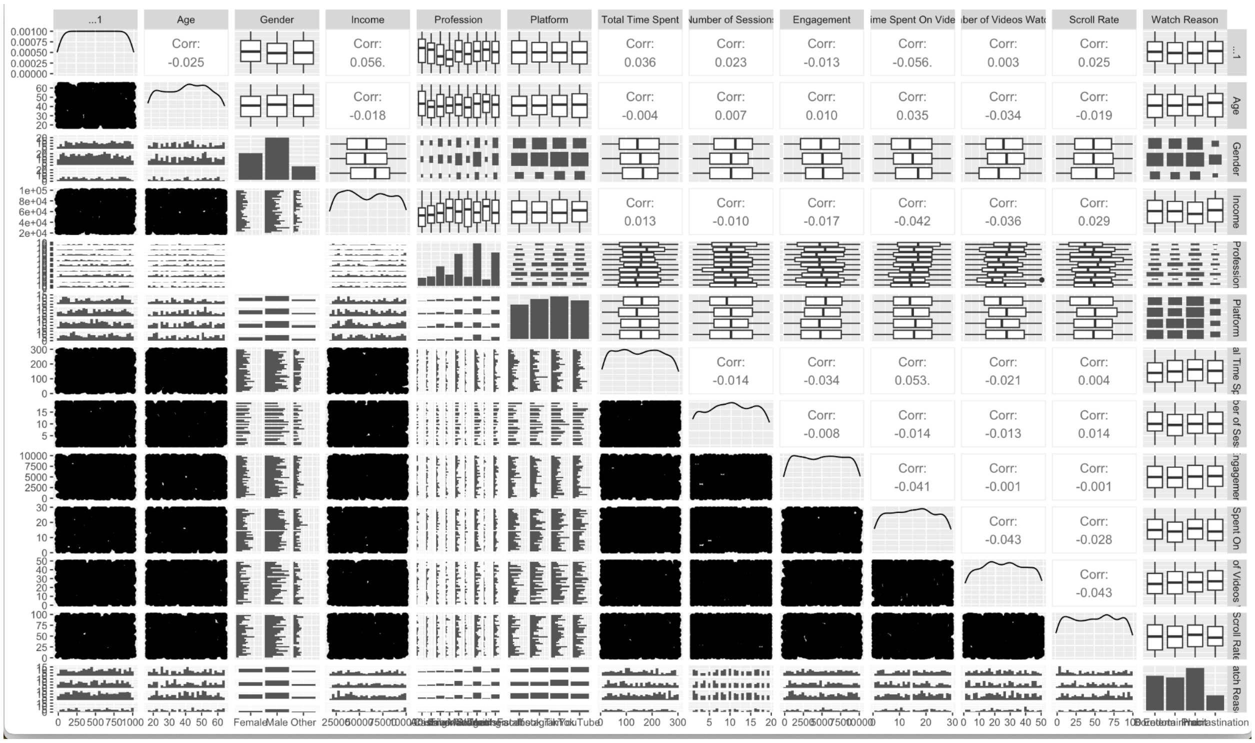
Up to this point, our statistical tests have not concluded any statistical significance. However, we will conduct further testing to see if there are any conclusions we can make on our research question.

CONTRIBUTIONS

Emily Zhao worked on writing introduction and background information and the analysis of the residuals. Derrick Chun generated scatter plots and residual plots. Linh Lai wrote and generated scatterplots. Jiakai Lin wrote. Andrew Gibson wrote and generated scatter plots and residual plots.

PART 4:

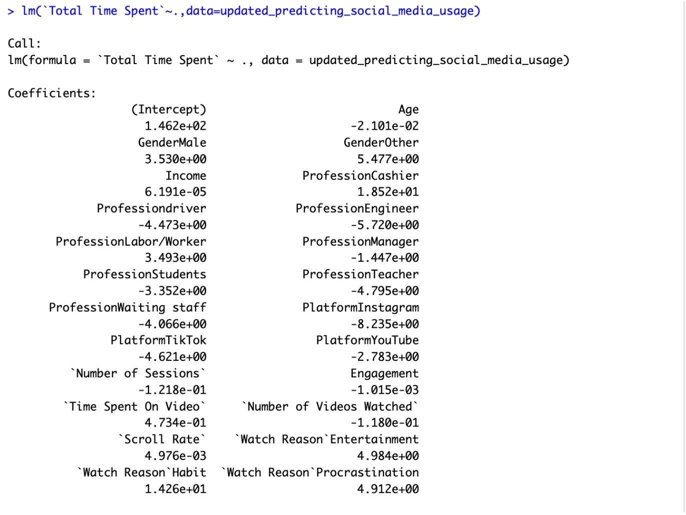
MULTICOLLINEARITY



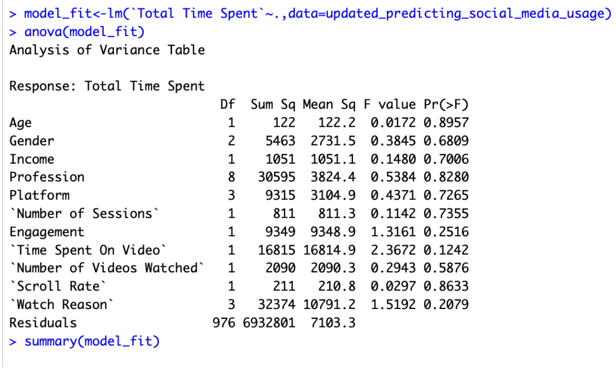
ANALYSIS

There are no strong correlations between any of the variables. There is no multicollinearity, or even predictor variables that are predictive of other predictor variables.

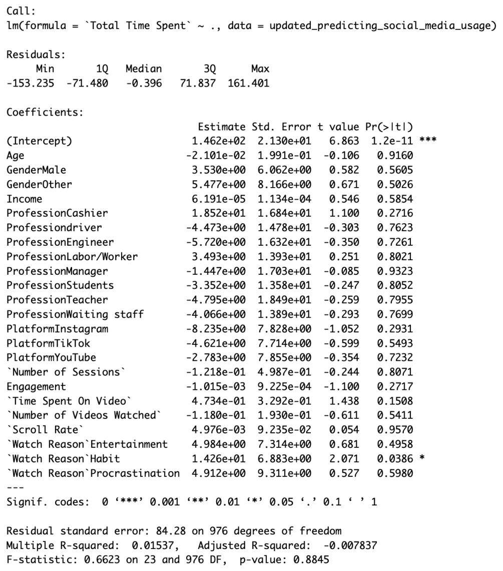
MODEL FIT



```
> lm(`Total Time Spent`~.,data=updated_predicting_social_media_usage)

Call:
lm(formula = `Total Time Spent` ~ ., data = updated_predicting_social_media_usage)

Coefficients:
                (Intercept)                               Age
                  1.462e+02                        -2.101e-02
                 GenderMale                       GenderOther
                  3.530e+00                         5.477e+00
                     Income                 ProfessionCashier
                  6.191e-05                         1.852e+01
           Professiondriver                ProfessionEngineer
                 -4.473e+00                        -5.720e+00
      ProfessionLabor/Worker                 ProfessionManager
                  3.493e+00                        -1.447e+00
          ProfessionStudents                 ProfessionTeacher
                 -3.352e+00                        -4.795e+00
    ProfessionWaiting staff                 PlatformInstagram
                 -4.066e+00                        -8.235e+00
             PlatformTikTok                  PlatformYouTube
                 -4.621e+00                        -2.783e+00
        `Number of Sessions`                        Engagement
                 -1.218e-01                        -1.015e-03
       `Time Spent On Video`     `Number of Videos Watched`
                  4.734e-01                        -1.180e-01
               `Scroll Rate`    `Watch Reason`Entertainment
                  4.976e-03                         4.984e+00
         `Watch Reason`Habit  `Watch Reason`Procrastination
                  1.426e+01                         4.912e+00
```



```
> model_fit<-lm(`Total Time Spent`~.,data=updated_predicting_social_media_usage)
> anova(model_fit)
Analysis of Variance Table

Response: Total Time Spent
                           Df  Sum Sq Mean Sq F value Pr(>F)
Age                         1     122   122.2  0.0172 0.8957
Gender                      2    5463  2731.5  0.3845 0.6809
Income                      1    1051  1051.1  0.1480 0.7006
Profession                  8   30595  3824.4  0.5384 0.8280
Platform                    3    9315  3104.9  0.4371 0.7265
`Number of Sessions`        1     811   811.3  0.1142 0.7355
Engagement                  1    9349  9348.9  1.3161 0.2516
`Time Spent On Video`       1   16815 16814.9  2.3672 0.1242
`Number of Videos Watched`  1    2090  2090.3  0.2943 0.5876
`Scroll Rate`               1     211   210.8  0.0297 0.8633
`Watch Reason`              3   32374 10791.2  1.5192 0.2079
Residuals                 976 6932801  7103.3
> summary(model_fit)
```



```
Call:
lm(formula = `Total Time Spent` ~ ., data = updated_predicting_social_media_usage)

Residuals:
     Min       1Q   Median       3Q      Max 
-153.235  -71.480   -0.396   71.837  161.401 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   1.462e+02  2.130e+01   6.863  1.2e-11 ***
Age                          -2.101e-02  1.991e-01  -0.106   0.9160    
GenderMale                    3.530e+00  6.062e+00   0.582   0.5605    
GenderOther                   5.477e+00  8.166e+00   0.671   0.5026    
Income                        6.191e-05  1.134e-04   0.546   0.5854    
ProfessionCashier             1.852e+01  1.684e+01   1.100   0.2716    
Professiondriver             -4.473e+00  1.478e+01  -0.303   0.7623    
ProfessionEngineer           -5.720e+00  1.632e+01  -0.350   0.7261    
ProfessionLabor/Worker        3.493e+00  1.393e+01   0.251   0.8021    
ProfessionManager            -1.447e+00  1.703e+01  -0.085   0.9323    
ProfessionStudents           -3.352e+00  1.358e+01  -0.247   0.8052    
ProfessionTeacher            -4.795e+00  1.849e+01  -0.259   0.7955    
ProfessionWaiting staff      -4.066e+00  1.389e+01  -0.293   0.7699    
PlatformInstagram            -8.235e+00  7.828e+00  -1.052   0.2931    
PlatformTikTok               -4.621e+00  7.714e+00  -0.599   0.5493    
PlatformYouTube              -2.783e+00  7.855e+00  -0.354   0.7232    
`Number of Sessions`         -1.218e-01  4.987e-01  -0.244   0.8071    
Engagement                   -1.015e-03  9.225e-04  -1.100   0.2717    
`Time Spent On Video`         4.734e-01  3.292e-01   1.438   0.1508    
`Number of Videos Watched`   -1.180e-01  1.930e-01  -0.611   0.5411    
`Scroll Rate`                 4.976e-03  9.235e-02   0.054   0.9570    
`Watch Reason`Entertainment   4.984e+00  7.314e+00   0.681   0.4958    
`Watch Reason`Habit           1.426e+01  6.883e+00   2.071   0.0386 *  
`Watch Reason`Procrastination 4.912e+00  9.311e+00   0.527   0.5980    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84.28 on 976 degrees of freedom
Multiple R-squared:  0.01537,	Adjusted R-squared:  -0.007837 
F-statistic: 0.6623 on 23 and 976 DF,  p-value: 0.8845
```

Adjusted R-squared: -0.007837
R-squared - 0.01537
H0: All variables are equal to 0, all x variables and y are independent
HA: At least one variable is not equal to 0, x and y are not independent

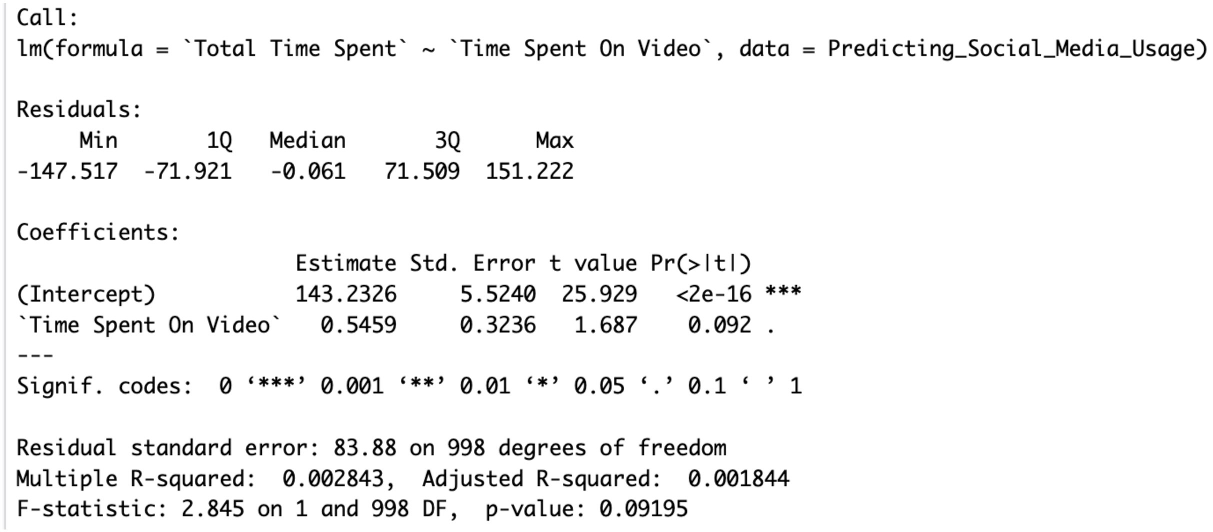
ANALYSIS

A low R-squared result and an even lower adjusted R-squared value really suggests that there is not any correlation between all the covariables and the time spent on social media. This is the adjusted R-squared value for the full model, but based on the T-values of the individual variables, we do not believe that removing any covariates would improve the R-squared score significantly. We cannot reject the null hypothesis. All variables are equal to zero and are not predictive of time spent on social media. This is not only because the R-squared value and the F-statistic are low, but also because individual T-tests would fail a hypothesis test for correlation.

VARIABLE SELECTION

The tests for every individual variable did not result in any of the null hypothesis being rejected, and graphs did not look correlated at all. Variable selection would be of no use, because there were no variables included in the model in the first place, so there is no point in trying stepwise selection or some other sort to remove “useless” variables, because there aren’t any variables in the first place. We do not believe that a transformation to the data would help either. However, we will try stepwise regression to see which of the variables are the “best” ones out of all of them.

STEPWISE REGRESSION



```
Call:
lm(formula = `Total Time Spent` ~ `Time Spent On Video`, data = Predicting_Social_Media_Usage)

Residuals:
     Min       1Q   Median       3Q      Max 
-147.517  -71.921   -0.061   71.509  151.222 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           143.2326     5.5240  25.929   <2e-16 ***
`Time Spent On Video`   0.5459     0.3236   1.687    0.092 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 83.88 on 998 degrees of freedom
Multiple R-squared:  0.002843,	Adjusted R-squared:  0.001844 
F-statistic: 2.845 on 1 and 998 DF,  p-value: 0.09195
```

Regression Model = Total Time Spent = $\beta_0 + \beta_1$ × Time Spent on Video

Intercept (β_0): 143.2326

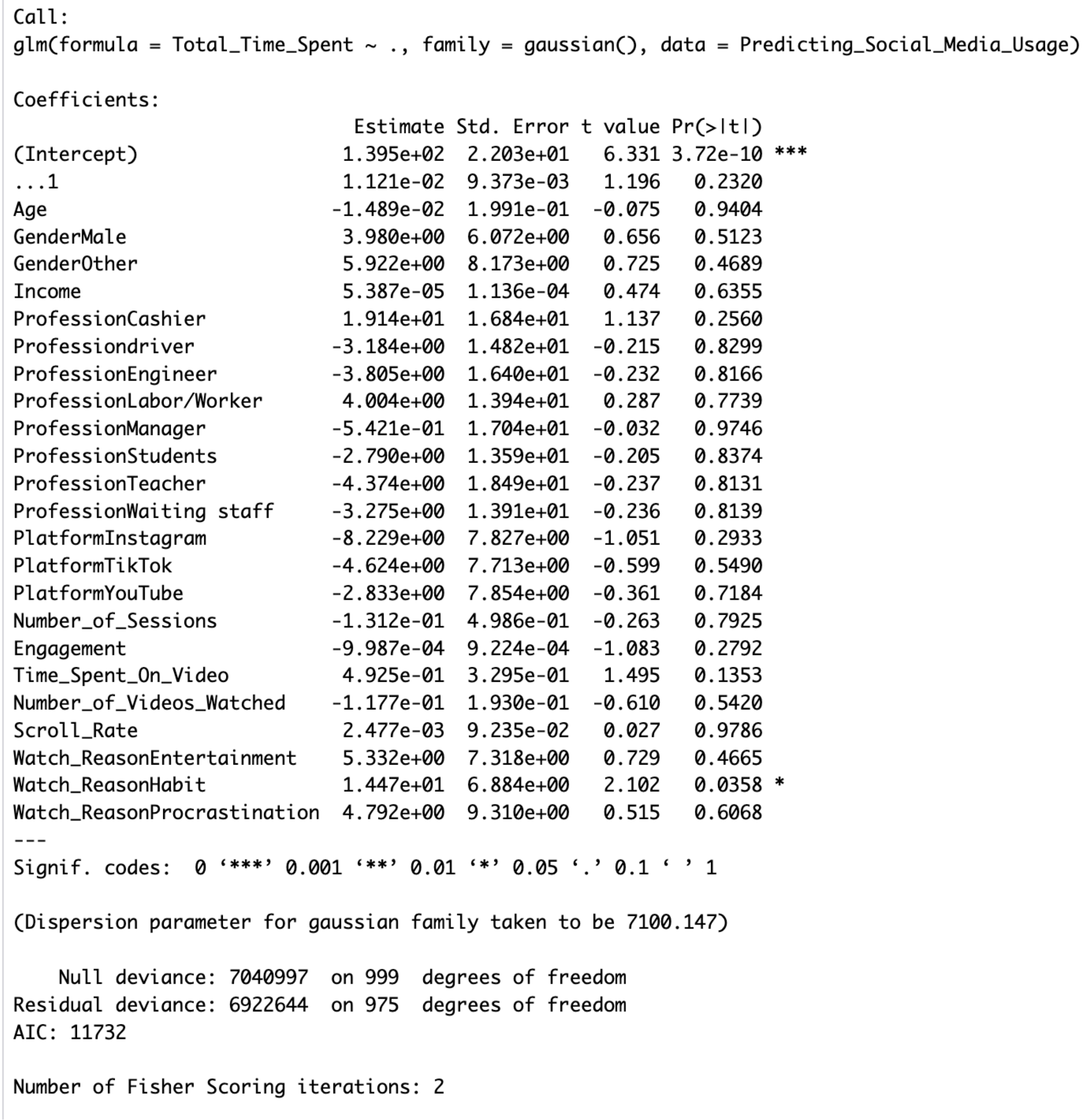
Coefficient for “Time Spent on Video” (β_1):: 0.5459

Interpretation: For every unit increase in minutes spent watching videos on social media, the total time spent on social media increases by approximately 0.5459 minutes (33 seconds).

ANALYSIS

This stepwise regression model predicts "Total Time Spent" on social media based on "Time Spent On Video." The intercept indicates when video-watching time is zero, the total time spent on social media is 143.2326. β_1, "Time Spent On Video", (0.5459) suggests a slight positive relationship but is not statistically significant (p-value = 0.092). We preferred a level of significance of 0.05, but this would be acceptable at 0.10. Despite not being statistically significant, it is the best predictor variable out of all of the others. The model explains only 0.28% of the variance in "Total Time Spent" (R-squared = 0.002843) is dependent on time spent on video. Overall, the results suggest that "Time Spent On Video" has a minimal impact on total time spent, and additional predictors are needed for better explanatory power. This is not what was expected, as we thought that the time spent on a video would increase time spent on social media, as videos are on social media, and would take

GENERALIZED LINEAR MODEL



```
Call:
glm(formula = Total_Time_Spent ~ ., family = gaussian(), data = Predicting_Social_Media_Usage)

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  1.395e+02  2.203e+01   6.331 3.72e-10 ***
...1                         1.121e-02  9.373e-03   1.196   0.2320    
Age                         -1.489e-02  1.991e-01  -0.075   0.9404    
GenderMale                   3.980e+00  6.072e+00   0.656   0.5123    
GenderOther                  5.922e+00  8.173e+00   0.725   0.4689    
Income                       5.387e-05  1.136e-04   0.474   0.6355    
ProfessionCashier            1.914e+01  1.684e+01   1.137   0.2560    
Professiondriver            -3.184e+00  1.482e+01  -0.215   0.8299    
ProfessionEngineer          -3.805e+00  1.640e+01  -0.232   0.8166    
ProfessionLabor/Worker       4.004e+00  1.394e+01   0.287   0.7739    
ProfessionManager           -5.421e-01  1.704e+01  -0.032   0.9746    
ProfessionStudents          -2.790e+00  1.359e+01  -0.205   0.8374    
ProfessionTeacher           -4.374e+00  1.849e+01  -0.237   0.8131    
ProfessionWaiting staff     -3.275e+00  1.391e+01  -0.236   0.8139    
PlatformInstagram           -8.229e+00  7.827e+00  -1.051   0.2933    
PlatformTikTok              -4.624e+00  7.713e+00  -0.599   0.5490    
PlatformYouTube             -2.833e+00  7.854e+00  -0.361   0.7184    
Number_of_Sessions          -1.312e-01  4.986e-01  -0.263   0.7925    
Engagement                  -9.987e-04  9.224e-04  -1.083   0.2792    
Time_Spent_On_Video          4.925e-01  3.295e-01   1.495   0.1353    
Number_of_Videos_Watched    -1.177e-01  1.930e-01  -0.610   0.5420    
Scroll_Rate                  2.477e-03  9.235e-02   0.027   0.9786    
Watch_ReasonEntertainment    5.332e+00  7.318e+00   0.729   0.4665    
Watch_ReasonHabit            1.447e+01  6.884e+00   2.102   0.0358 *  
Watch_ReasonProcrastination  4.792e+00  9.310e+00   0.515   0.6068    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 7100.147)

    Null deviance: 7040997  on 999  degrees of freedom
Residual deviance: 6922644  on 975  degrees of freedom
AIC: 11732

Number of Fisher Scoring iterations: 2
```

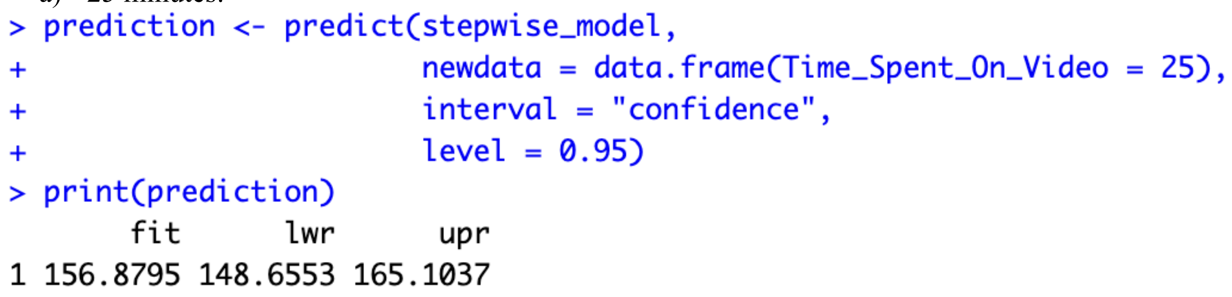
note: the response variable is continuous

ANALYSIS

The output summarizes a Generalized Linear Model (GLM) predicting "Total_Time_Spent" on social media using various predictors where the response variable is continuous. Almost all predictors don’t seem to significantly impact the total time spent since their p-value is greater than 0.05. The only significant predictor is "Watch_ReasonHabit" (p-value = 0.0358), indicating that habitual use positively and significantly increases time spent on social media. However, the other watch reasons were not significant, which is why that qualitative variable was not included in the stepwise regression.

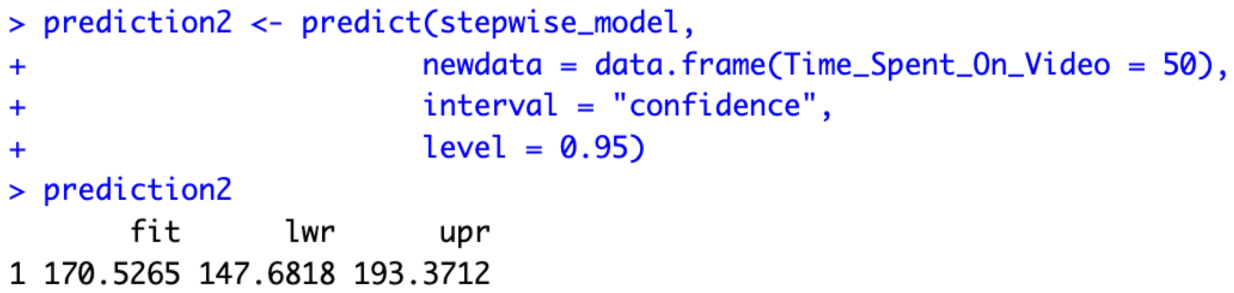
CONFIDENCE AND PREDICTION INTERVALS

a) 25 minutes:



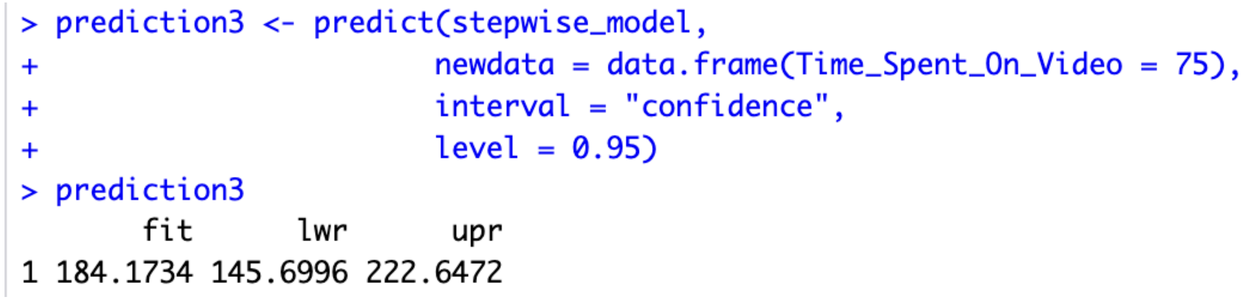
```
> prediction <- predict(stepwise_model,
+                           newdata = data.frame(Time_Spent_On_Video = 25),
+                           interval = "confidence",
+                           level = 0.95)
> print(prediction)
       fit      lwr      upr
1 156.8795 148.6553 165.1037
```

b) 50 minutes:



```
> prediction2 <- predict(stepwise_model,
+                           newdata = data.frame(Time_Spent_On_Video = 50),
+                           interval = "confidence",
+                           level = 0.95)
> prediction2
       fit      lwr      upr
1 170.5265 147.6818 193.3712
```

c) 75 minutes:



```
> prediction3 <- predict(stepwise_model,
+                           newdata = data.frame(Time_Spent_On_Video = 75),
+                           interval = "confidence",
+                           level = 0.95)
> prediction3
       fit      lwr      upr
1 184.1734 145.6996 222.6472
```

ANALYSIS

How does the total time spent on social media (Total_Time_Spent) depend on the time spent watching video (Time_Spent_On_Video)?

Regression Model: Total Time Spent = β_0 + β_1 × Time Spent on Video (25, 50, 75)

β_0 = 143.2326

β_1 = 0.5459

25 minutes spent on videos: (148.655, 165.103):

For 25 minutes spent on video, the predicted total time spent on social media is between 148.655 minutes and 165.103 minutes. We are 95% confident the individuals who spend 25 minutes on videos will spend approximately between 148.66 minutes and 165.1 minutes in total on social media.

50 minutes spent on videos: (147.682, 193.371):

For 50 minutes spent on video, the predicted total time spent on social media is between 147.682 and 193.371 minutes. We are 95% confident the individuals who spend 50 minutes on videos will spend approximately between 147.682 minutes and 193.371 minutes in total on social media.

75 minutes spent on videos: (145.7, 222.647):

For 75 minutes spent on video, the predicted total time spent on social media is between 1145.7 and 222.647 minutes. We are 95% confident the individuals who spend 75 minutes on videos will spend approximately between 145.7 minutes and 222.647 minutes in total on social media.

With 95% confidence we found three predicted intervals. The intervals are for the total minutes spent on social media; because zero is not included in the intervals, we can conclude the results are statistically significant, and people spending time on videos contribute positively to the total time on social media. Moreover, the positive slope suggests that increasing time spent on social media is correlated to increase in total time on social media. Thus, the significant β_1 in our model indicates that the "time spent on video" is a significant predictor to predict social media usage of individuals.

SUMMARY

Our model indicates that "Watch Reason Habit" is the most essential predictor of the response variable "Total Time Spent" on social media. However, we cannot include all watch reasons in the model because the other factor levels of watch reason are not significant predictors. The only variable that we will include is time spent on video because it was the only variable selected by the stepwise procedure. Although the p-value for time spent on videos is not satisfied at a 0.05 level of significance, we can still reject the null hypothesis at a level of significance of 0.10. This means that there is sufficient evidence to suggest that minutes spent on videos is predictive of total time spent on social media. The other variables are excluded because we failed to reject the null hypothesis for them, and there is not enough evidence to suggest they are predictive of total time spent. We even are able to see that they are not predictive just by looking at the scatterplots; the data is randomly scattered, and would not be worth checking in the first place.

CONTRIBUTIONS

Linh Lai interpreted the R output and also tested individual variables' linear relationships. Derrick Chun visualized and analyzed the dataset's multicollinearity, model fit, stepwise regression, generalized linear model, confidence and prediction intervals. Emily Zhao interpreted the results of multicollinearity, linear regression, and stepwise regression. Andrew Gibson wrote analysis and formatted the final report document. Jiakai Lin wrote analysis and assisted with model fitting.

SOURCES

https://www.kaggle.com/datasets/zeesolver/dark-web
R + RStudio