Matching Analysis

The dataset D5.2 (described in C5.2) contains information of online news articles published by Mashable (www.mashable.com). Some of these articles contains videos.

- **Our primary question in this assignment is whether including at least one video in an article leads to the article being shared more in social media.**
- **Accordingly, the key outcome is "shares"—the number of social media shares for each article.**
- **For tasks 1-3, the treatment indicator will a variable that equals 1 if the number of videos included in an article (num_videos) is non-zero, and which equals 0 otherwise.**

To prepare for task 1-3, I create binomial variable "treatment" with this condition:

- The number of videos included in an article (num_videos) is non-zero 0: 1
- The number of videos included in an article (num_videos) is zero: 0

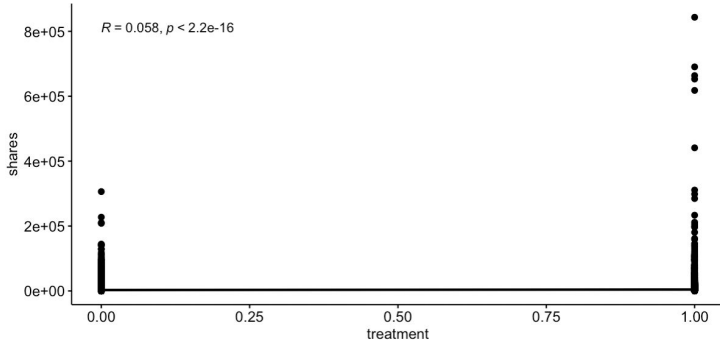
```
# Convert num_videos to numeric and create treatment variable
data$num_videos <- as.numeric(as.character(data$num_videos))
data$treatment <- ifelse(data$num_videos > 0, 1, 0)
```

Task 1

Based on linear regression results, is the treatment associated to a typically larger or lower number of shares?

```
        Pearson's product-moment correlation

data:  data$treatment and data$shares
t = 11.464, df = 38710, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.04823390 0.06808963
sample estimates:
       cor
0.05816752
```

Running correlation test between treatment (video in an article) and shares:
Results indicate a significantly moderate, positive relationship between article with videos and number of shares r(38710) = 0.05, p < 0.01.




```
> # Fit a linear regression model
> model <- lm(shares ~ treatment, data = data)
> # Print the model summary
> summary(model)

Call:
lm(formula = shares ~ treatment, data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -4309  -2310  -1691   -491 838990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2891.47      73.58   39.30   <2e-16 ***
treatment    1418.71     123.76   11.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11640 on 38710 degrees of freedom
Multiple R-squared:  0.003383,  Adjusted R-squared:  0.003358
F-statistic: 131.4 on 1 and 38710 DF,  p-value: < 2.2e-16
```

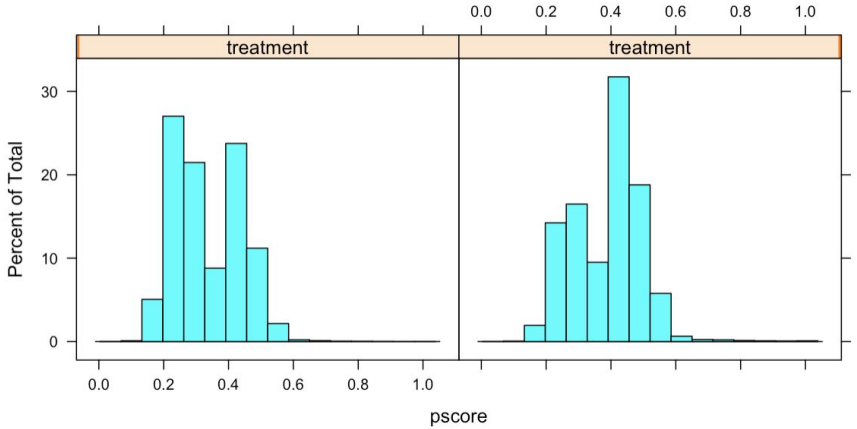
Running linear regression on treatment and shares:
Based on the given output, the coefficient estimate for treatment = 1418.71, with a standard error of 123.76, p-value < 0.05. This is a positive and statistically significant relationship between articles with videos and the number of shares, indicates that the number of videos included in an article is associated with a typically larger number of shares.

Specifically, having videos tends to be associated with about 1418.71 point increase (49% increase) in the number of shares.

Task 2

a. Evaluate the propensity score overlap between treated and non-treated subsamples.

```
library(tableone)
library(MatchIt)

# Convert categorical variables to factors
data$category <- as.factor(data$category)
data$weekday <- as.factor(data$weekday)

# Check balance of covariates between treated and untreated groups
print(CreateTableOne(vars = c("num_imgs", "num_keywords", "category", "weekday",
"shares"), data = data, strata = "treatment"), smd = TRUE)

# Checking overlap
# Fit logistic regression model to calculate propensity scores
data$pscore <- predict(glm(treatment ~ num_imgs + num_keywords + category +
weekday + shares, data = data, family = "binomial"), type = "response")
# Plot histograms of propensity scores for treated and untreated groups
histogram(~ pscore | treatment, data = data)
```



- The majority of observations have a pscore between 0.1 and 0.6 and distributions are overall similar. This suggests that the covariate balance between the groups is satisfactory.
- The distributions don't overlap in the range of pscore above 0.6. This suggests that there may be some covariate imbalance in this range, which could affect the reliability of the treatment effect estimate.

Task 2

b. Create a matched sample based on logistic propensity scores and in a way that accounts for overlap considerations

```
# Perform matching
matched <- matchit(treatment ~ num_imgs + num_keywords + category + weekday + shares, method = "nearest", data = data)
# Create matched data set
data_matched = match.data(matched)
data_matched = data_matched[data_matched$pscore<=0.6,] # focus on area with enough overlap
dim(data_matched)

# Check balance of covariates between matched treated and untreated groups
data_matched <- match.data(matched)
print(CreateTableOne(vars = c("num_imgs", "num_keywords", "category", "weekday", "shares"), data = data_matched, strata = "treatment"), smd = TRUE)
```

```
> dim(data_matched)
[1] 27115    42
```

About 27,000 observations are put in the matched sample

Task 2

c. Assess the matched sample in terms of covariate balancing. In your judgement, has the matching procedure been successful?

```
> # Check balance of covariates between treated and untreated groups
> print(CreateTableOne(vars = c("num_imgs", "num_keywords", "category", "weekday", "shares"), data = data, strata =
"treatment"), smd = TRUE)
                         Stratified by treatment
                          0                 1                 p      test SMD
  n                         25026             13686
  num_imgs (mean (SD))       4.68 (8.12)       4.20 (8.42)     <0.001     0.059
  num_keywords (mean (SD))   7.19 (1.94)       7.30 (1.84)     <0.001     0.057
  category (%)                                                 <0.001     0.441
     business                4893 (19.6)       1365 (10.0)
     entertainment           3152 (12.6)       2973 (21.7)
     lifestyle               1631 ( 6.5)        468 ( 3.4)
     socialmedia             1642 ( 6.6)        681 ( 5.0)
     tech                    5275 (21.1)       2071 (15.1)
     world                   8433 (33.7)       6128 (44.8)
  weekday (%)                                                  <0.001     0.077
     friday                  3561 (14.2)       2008 (14.7)
     monday                  4164 (16.6)       2313 (16.9)
     saturday                1677 ( 6.7)        724 ( 5.3)
     sunday                  1791 ( 7.2)        853 ( 6.2)
     thursday                4619 (18.5)       2505 (18.3)
     tuesday                 4541 (18.1)       2683 (19.6)
     wednesday               4673 (18.7)       2600 (19.0)
  shares (mean (SD))      2891.47 (6524.88) 4310.18 (17476.82) <0.001     0.108
```

Unmatched

```
> # Check balance of covariates between matched treated and untreated groups
> data_matched <- match.data(matched)
> print(CreateTableOne(vars = c("num_imgs", "num_keywords", "category", "weekday", "shares"), data = data_matched,
strata = "treatment"), smd = TRUE)
                         Stratified by treatment
                          0                 1                 p      test SMD
  n                         13686             13686
  num_imgs (mean (SD))       4.31 (7.36)       4.20 (8.42)     0.270     0.013
  num_keywords (mean (SD))   7.33 (1.91)       7.30 (1.84)     0.087     0.021
  category (%)                                                 0.209     0.032
     business                1253 ( 9.2)       1365 (10.0)
     entertainment           2941 (21.5)       2973 (21.7)
     lifestyle                491 ( 3.6)        468 ( 3.4)
     socialmedia              676 ( 4.9)        681 ( 5.0)
     tech                    2149 (15.7)       2071 (15.1)
     world                   6176 (45.1)       6128 (44.8)
  weekday (%)                                                  0.842     0.020
     friday                  2018 (14.7)       2008 (14.7)
     monday                  2243 (16.4)       2313 (16.9)
     saturday                 771 ( 5.6)        724 ( 5.3)
     sunday                   843 ( 6.2)        853 ( 6.2)
     thursday                2498 (18.3)       2505 (18.3)
     tuesday                 2700 (19.7)       2683 (19.6)
     wednesday               2613 (19.1)       2600 (19.0)
  shares (mean (SD))      3304.59 (8354.53) 4310.18 (17476.82) <0.001     0.073
```

Matched

- For the unmatched, the untreated sample contains about 25,000 observations while the treated one only around 13,700 observations. For the matched sample, the treated and untreated samples are the same size (13,686 observations) and the sample is well balanced. This result suggests that matching procedure was successful at recreating the parallel worlds situation and we can treat our data largely as if it were from a randomized controlled experiments
- SMDs for covariates are below the conventional threshold of 0.1, indicating that there are no substantial differences between the two groups in terms of the covariates included in the matching procedure

Task 3

a. Based on your analysis above, provide a matching ATE estimate. Do videos increase the number of shares? By how much? For simplicity, base your answer on a regression of the outcome on the treatment indicator (ie, do not include other covariates).

```
> summary(lm(shares ~ treatment, data = data_matched))

Call:
lm(formula = shares ~ treatment, data = data_matched)

Residuals:
   Min     1Q Median     3Q    Max
 -4309  -2869  -2205   -810 838990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3304.6      117.1  28.224  < 2e-16 ***
treatment     1005.6      165.6   6.073 1.27e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13700 on 27370 degrees of freedom
Multiple R-squared:  0.001346,  Adjusted R-squared:  0.001309
F-statistic: 36.88 on 1 and 27370 DF,  p-value: 1.272e-09
```

Matched

```
> summary(lm(shares ~ treatment, data = data)) # compare to estimate from the unmatched sample

Call:
lm(formula = shares ~ treatment, data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -4309  -2310  -1691   -491 838990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2891.47      73.58   39.30   <2e-16 ***
treatment    1418.71     123.76   11.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11640 on 38710 degrees of freedom
Multiple R-squared:  0.003383,  Adjusted R-squared:  0.003358
F-statistic: 131.4 on 1 and 38710 DF,  p-value: < 2.2e-16
```

Unmatched

- Matched: ATE = 1005.6, SE = 165.6, p-value < 0.01. The estimate is statistically significant at a high level of confidence. This means that having videos tends to be associated with about a 1005.6 point increase (30.4% increase) in the number of shares per day, after controlling for other covariates in the model.
- Unmatched: ATE = 1418.71, SE = 132.76, p-value < 0.01. The estimate is also statistically significant at a high level of confidence. This means that having videos tends to be associated with about a 1418.71 point increase (49% increase) in the number of shares per day, after controlling for other covariates in the model.
- The ATE estimate from the matched data is lower than the ATE estimate from the unmatched data, suggesting that the matching procedure has reduced the bias in the estimation of the treatment effect. However, the ATE estimate from the unmatched data is still statistically significant and suggests a strong positive effect of having videos on the number of shares.

Task 3

b. Provide a rationale that explains the sign of the difference between the estimate of 3.a and 1.a (i.e., your rationale must describe some form of behavior for why one estimate is larger than the other).

- The difference between the estimates in 3.a and 1.a can be explained by the fact that the two analyses use different methods to estimate the treatment effect and control for confounding variables.
- In task 1.a, a linear regression model is used to estimate the association between the treatment (i.e. videos) and the outcome variable shares, without controlling for other covariates. This analysis does not account for the potential influence of confounding variables that may affect the relationship between treatment and outcome.
- In contrast, in task 3.a, a matching procedure is used to balance the covariate distributions between the treated and untreated groups, and a linear regression model is used to estimate the ATE of treatment on shares, after controlling for other covariates. This analysis aims to reduce the influence of confounding variables and provide a more accurate estimate of the treatment effect.
- The difference in the estimated treatment effect between the two analyses may be due to the presence of confounding variables in the data. The correlation analysis in task 1.a may not have fully accounted for the influence of these variables, leading to a larger estimated treatment effect. However, the matching procedure in task 3.a may have successfully balanced the covariate distributions and reduced the influence of these variables, resulting in a lower estimated treatment effect.
- Therefore, the sign of the difference between the estimates in 3.a and 1.a may be due to the presence of confounding variables and the effectiveness of the matching procedure in reducing their influence on the treatment effect estimate.

c. Suppose that the unconfoundedness assumption holds: what could then be the “fudge factor” (discussed in class) in this case? Explain.

- Measurement error in the outcome variable (i.e. shares): If the outcome variable is measured with error, then the estimated treatment effect may be biased due to incorrect measurement of the outcome.
- Omitted variable bias: Important covariates that affect the relationship between treatment and outcome are not included in the analysis. This can lead to bias in the treatment effect estimate.

Task 4

Propose a propensity score matching analysis that addresses the question of whether the specific number of videos included in an article (not whether a video is included) has a causal effect on the number of shares. For this analysis, you will need to make decisions about: (i) how to define the treatment indicator, (ii) what subsample of the full dataset to use. Note: there is no one single correct analysis.

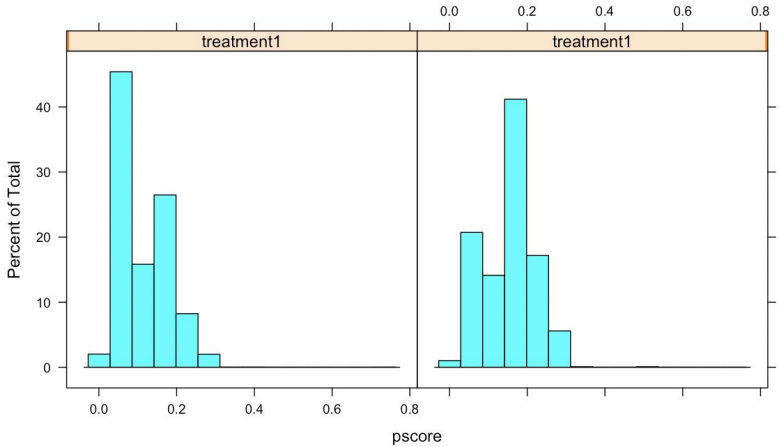
Given that the primary question in this assignment is whether including at least one video in an article leads to the article being shared more in social media, for task 4, I decided to check whether having at least 1 video in an article has a causal effect on the number of shares. For the subset, I choose random 10,000 observations from the dataset.

```
########### task 4
# Select a subsample of the full dataset
set.seed(123)
data_subsample <- data[sample(nrow(data), 10000), ]

# Create treatment variable, articles included > 1 video = 1, < 1 video = 0
data_subsample$treatment1 <- ifelse(data_subsample$num_videos > 1, 1, 0)
```

Then I perform similar steps like the last 2 task:

```
# Convert categorical variables to factors
data_subsample$category <- as.factor(data_subsample$category)
data_subsample$weekday <- as.factor(data_subsample$weekday)

# Check balance of covariates between treated and untreated groups
print(CreateTableOne(vars = c("num_imgs", "num_keywords", "category", "weekday", "shares"), data = data_subsample, strata = "treatment"), smd =
TRUE)
```

```
# Checking overlap
# Fit logistic regression model to calculate propensity scores
data_subsample$pscore <- predict(glm(treatment1 ~ num_imgs + num_keywords + category + weekday + shares, data = data_subsample, family =
"binomial"), type = "response")
# Plot histograms of propensity scores for treated and untreated groups
histogram(~ pscore | treatment1, data = data_subsample)
```



The sub-sample got me the same results:

- The majority of observations have a pscore between 0.1 and 0.6 and distributions are overall similar. This suggests that the covariate balance between the groups is satisfactory.
- The distributions don't overlap in the range of pscore above 0.6. This suggests that there may be some covariate imbalance in this range, which could affect the reliability of the treatment effect estimate.

Then, I perform match with pscore 0.6:

```
# Perform matching
matched1 <- matchit(treatment1 ~ pscore, method = "nearest", data = data_subsample)
# Create matched data set
data_matched1 = match.data(matched1)
data_matched1 = data_matched1[data_matched1$pscore<=0.6,] # focus on area with enough overlap
dim(data_matched1)
```

I did the same step with checking balance:

- SMD values for num_imgs, num_keywords, category, weekday, and shares are below the recommended threshold of 0.1, indicating good balance

```
> # Check balance of covariates between matched treated and untreated groups
> print(CreateTableOne(vars = c("num_imgs", "num_keywords", "category", "weekday", "shares"), data = data_matched1, strata = "tr
eatment"), smd = TRUE)
                        Stratified by treatment
                         0                 1                 p      test SMD
  n                        791               1573
  num_imgs (mean (SD))     3.47 (6.39)       2.96 (6.33)       0.068      0.079
  num_keywords (mean (SD)) 7.27 (1.89)       7.31 (1.85)       0.577      0.024
  category (%)                                                 0.063      0.139
     business              57 ( 7.2)         101 ( 6.4)
     entertainment        200 (25.3)         459 (29.2)
     lifestyle             27 ( 3.4)          45 ( 2.9)
     socialmedia           77 ( 9.7)         106 ( 6.7)
     tech                  86 (10.9)         159 (10.1)
     world                344 (43.5)         703 (44.7)
  weekday (%)                                                  0.362      0.110
     friday               107 (13.5)         221 (14.0)
     monday               141 (17.8)         263 (16.7)
     saturday              44 ( 5.6)          84 ( 5.3)
     sunday                66 ( 8.3)          93 ( 5.9)
     thursday             146 (18.5)         291 (18.5)
     tuesday              152 (19.2)         332 (21.1)
     wednesday            135 (17.1)         289 (18.4)
  shares (mean (SD))       2979.71 (6729.30) 4531.95 (21045.41) 0.043      0.099
```

Estimate treatment effect

```
> summary(lm(shares ~ treatment1, data = matched_data1))

Call:
lm(formula = shares ~ treatment1, data = matched_data1)

Residuals:
   Min     1Q Median     3Q    Max
 -4314  -3042  -2483  -1083 686058

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3683.2      512.4   7.188 8.78e-13 ***
treatment1     658.8      724.7   0.909    0.363
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17620 on 2362 degrees of freedom
Multiple R-squared:  0.0003498, Adjusted R-squared:  -7.342e-05
F-statistic: 0.8265 on 1 and 2362 DF,  p-value: 0.3634
```

p-value > 0.05, indicate that there is no evidence to suggest that having more than 1 video in an article has a causal effect on the number of shares

Conclude: The difference in the new treatment can lead to different effects. It's possible that the presence of any video in an article has a positive effect on the number of shares (task 1-3), but having more than one video (task 4) doesn't have an additional effect on shares. Additionally, the size of the subsample is much smaller than the full dataset, which may lead to less statistical power and higher uncertainty in the treatment effect estimate.