Project

Rishabh Kumar Kandoi | Aradhya Mathur | Ayush Singla | Richa Yadav

2022-12-07

Q1. The marketing team first wants to understand how many ads they are running on each social media platform, as well as the demographics of each social media platform's user base.

a. Create a relative frequency table and a corresponding relative frequency barplot to visualize the fraction of ads on each platform. Make sure to label the plot (title, axes), and comment on trends you observe

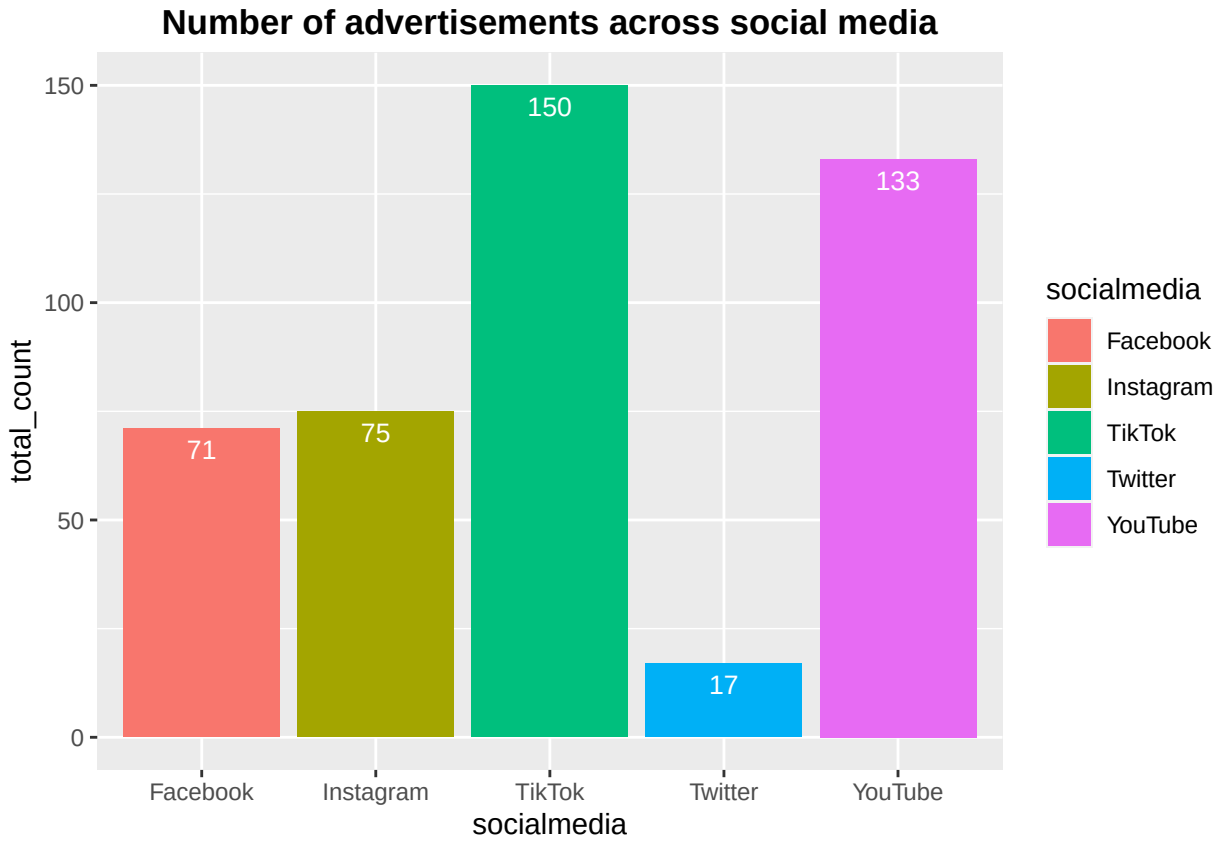
b. The CFO's ad strategy is supposed to run 10% of all ads on Twitter, 10% on Facebook, 20% on Instagram, 30% on TikTok, and 30% on YouTube. Is the marketing department following this strategy? Run an appropriate statistical test at the alpha = 0.05 significance level and comment on the results.

c. For each social media platform, calculate the variance, standard deviation, coefficient of variation, and skew of age, and visualize the distribution of age using an appropriate tool from descriptive statistics. Comment on any trends you see.

```
data = read.csv("ads4.csv")

# Part a)
grouped_data <- data %>%
    group_by(socialmedia) %>%
    summarise(total_count = n(), .groups = "drop") %>%
    as.data.frame()
grouped_data
```

```
##   socialmedia total_count
## 1    Facebook          71
## 2   Instagram          75
## 3      TikTok         150
## 4     Twitter          17
## 5     YouTube         133
```

```
ggplot(data = grouped_data, aes(x = socialmedia, y = total_count, fill = socialmedia)) + geom_bar(stat =
    geom_text(aes(label = total_count), vjust = 1.6, color = "white", position = position_dodge(0.9),
        size = 3.5) + ggtitle("Number of advertisements across social media") + theme(plot.title = eleme
    vjust = 0.5, face = "bold"))
```




```
paste(" Insight : More than 60% of the ads are running on tiktok and youtube with only 4% on twitter")
```

```
## [1] " Insight : More than 60% of the ads are running on tiktok and youtube with only 4% on twitter"
```

```
# Part b) H0 - Proportion Values: Twitter=0.1, Facebook=0.1, Instagram=0.2, TikTok=0.3,
# YouTube=0.3 H1 - Atleast one of these proportions does not hold

observed_exp_table = data.frame(Row = c("Twitter", "Facebook", "Instagram", "TikTok", "YouTube"),
    Observed = c("17", "71", "75", "150", "133"), Expected = c("44.6", "44.6", "89.2", "133.8",
        "133.8"))
observed_exp_table
```

```
##         Row Observed Expected
## 1   Twitter       17     44.6
## 2  Facebook       71     44.6
## 3 Instagram       75     89.2
## 4    TikTok      150    133.8
## 5   YouTube      133    133.8
```

```
chisq.test(c(17, 71, 75, 150, 133), p = c(0.1, 0.1, 0.2, 0.3, 0.3))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(17, 71, 75, 150, 133)
## X-squared = 36.933, df = 4, p-value = 1.859e-07
```

```
paste("P-value = 1.859e-07, Thus atleast one of the proportions does not hold. Insight: The observed an
```

[1] "P-value = 1.859e-07, Thus atleast one of the proportions does not hold. Insight: The observed and

expected distribution of ads across various social media platforms are not the same. Hence, marketing team needs to make necessary changes in order to follow their plans"

```
# Part c)
twitter_data = filter(data, socialmedia == "Twitter")["age"]
instagram_data = filter(data, socialmedia == "Instagram")["age"]
facebook_data = filter(data, socialmedia == "Facebook")["age"]
tiktok_data = filter(data, socialmedia == "TikTok")["age"]
youtube_data = filter(data, socialmedia == "YouTube")["age"]

paste("For Twitter Data: Var =", var(twitter_data), ", Standard Deviation =", sd(as.numeric(unlist(twit
    ", Mean =", mean(as.numeric(unlist(twitter_data))), ", Coefficient of Variation =", sd(as.numeric(u
    ", Skewness =", skewness(twitter_data))
```

[1] "For Twitter Data: Var = 14.7205882352941 , Standard Deviation = 3.83674187759538 , Mean = 38.7058823529412 , Coefficient of Variation = 0.09912555002906 , Skewness = 0.53297699232656"

```
paste("For Instagram Data: Var =", var(instagram_data), ", Standard Deviation =", sd(as.numeric(unlist(
    ", Mean =", mean(as.numeric(unlist(instagram_data))), ", Coefficient of Variation =", sd(as.numeric
    ", Skewness =", skewness(instagram_data))
```

[1] "For Instagram Data: Var = 27.8367567567568 , Standard Deviation = 5.27605503731308 , Mean = 27.12 , Coefficient of Variation = 0.194544802260807 , Skewness = 0.425110148502713"

```
paste("For Facebook Data: Var =", var(facebook_data), ", Standard Deviation =", sd(as.numeric(unlist(fa
    ", Mean =", mean(as.numeric(unlist(facebook_data))), ", Coefficient of Variation =", sd(as.numeric(
    ", Skewness =", skewness(facebook_data))
```

[1] "For Facebook Data: Var = 83.7106639839034 , Standard Deviation = 9.14935320030347 , Mean = 30.5070422535211 , Coefficient of Variation = 0.299909546270335 , Skewness = 0.184380814791779"

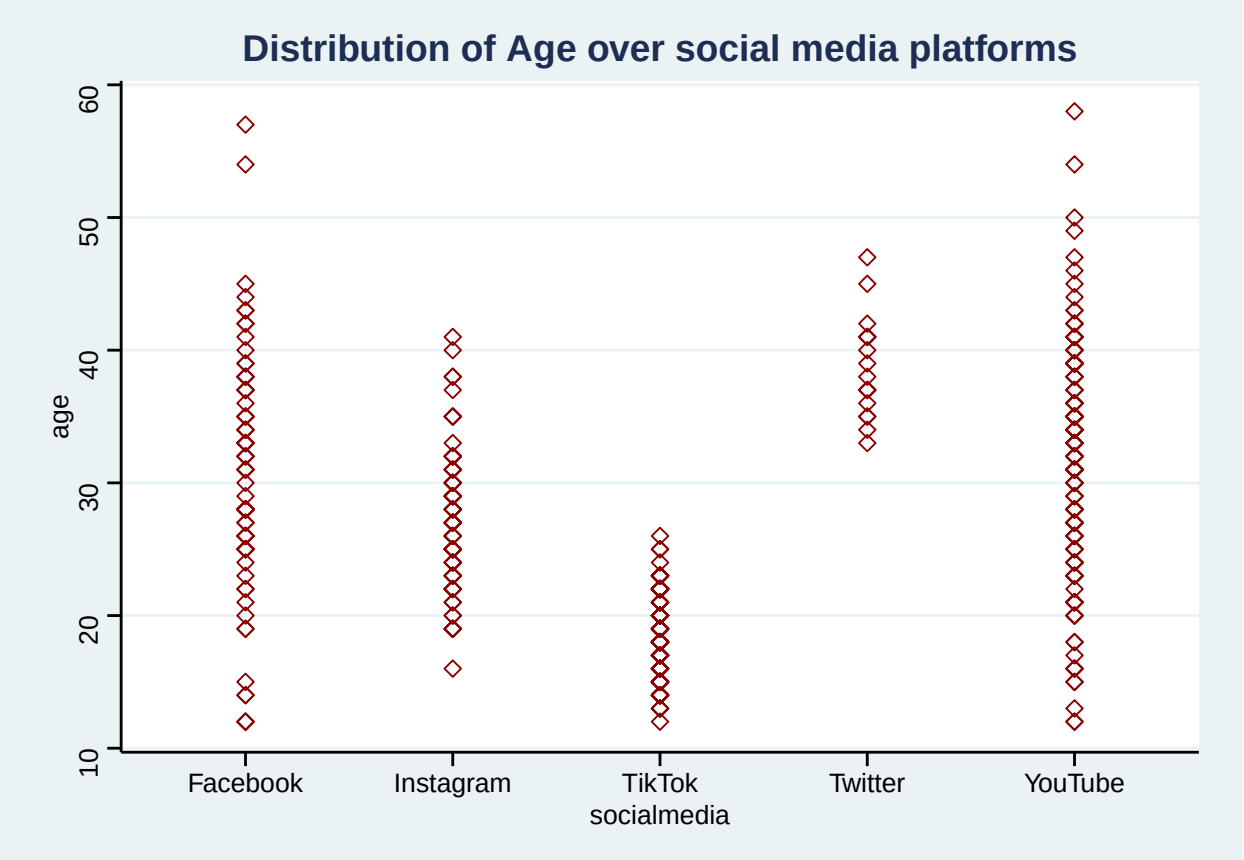
```
paste("For TikTok Data: Var =", var(tiktok_data), ", Standard Deviation =", sd(as.numeric(unlist(tiktok
    ", Mean =", mean(as.numeric(unlist(tiktok_data))), ", Coefficient of Variation =", sd(as.numeric(un
    ", Skewness =", skewness(tiktok_data))
```

[1] "For TikTok Data: Var = 9.41659955257271 , Standard Deviation = 3.0686478378225 , Mean = 18.4466666666667 , Coefficient of Variation = 0.166352430673428 , Skewness = 0.10988037076095"

```
paste("For YouTube Data: Var =", var(youtube_data), ", Standard Deviation =", sd(as.numeric(unlist(youtu
    ", Mean =", mean(as.numeric(unlist(youtube_data))), ", Coefficient of Variation =", sd(as.numeric(u
    ", Skewness =", skewness(youtube_data))
```

[1] "For YouTube Data: Var = 75.8395989974937 , Standard Deviation = 8.70859339948156 , Mean = 31.4436090225564 , Coefficient of Variation = 0.276959091853431 , Skewness = 0.0628277838423192"

```
ggplot(data, aes(x = socialmedia, y = age)) + geom_point(size = 2, shape = 23, color = "darkred") +
    theme_stata() + scale_color_stata() + ggtitle("Distribution of Age over social media platforms") +
    theme(plot.title = element_text(hjust = 0.5, vjust = 0.5, face = "bold"))
```




```
paste("Insight: Distribution of age across each social media platform looks normal")
```

[1] "Insight: Distribution of age across each social media platform looks normal"

Q2. The CEO of the company believes that ads differ in effectiveness (measured in terms of profit, or ad revenue - ad cost) depending on the season. However, is his intuition correct?

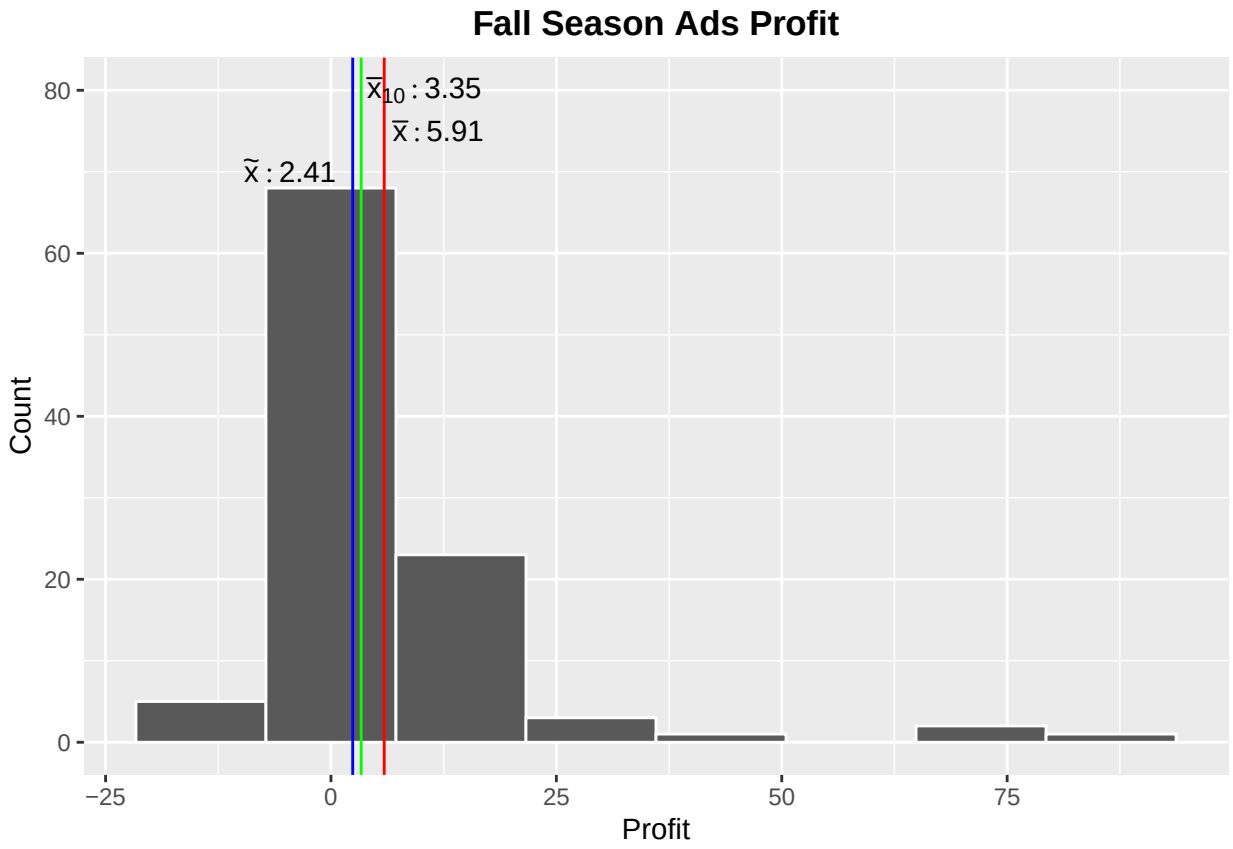
a. Visualize the data using four histograms (one from each season). On each plot, draw and label vertical lines for the mean, median, and 10% trimmed mean. Make sure to label the plots (title, axes, legend), and comment on trends you observe.

b. In particular, the CEO believes that summer ads yield more profit than winter ads. At the alpha = 0.05 significance level, run an appropriate statistical test (or series of tests) and comment on your results.

c. What if you wanted to compare all seasons at once at the alpha = 0.05 significance level with a familywise error of alphaFW E = 0.05? Run an appropriate test (or series of tests) and comment on your results.

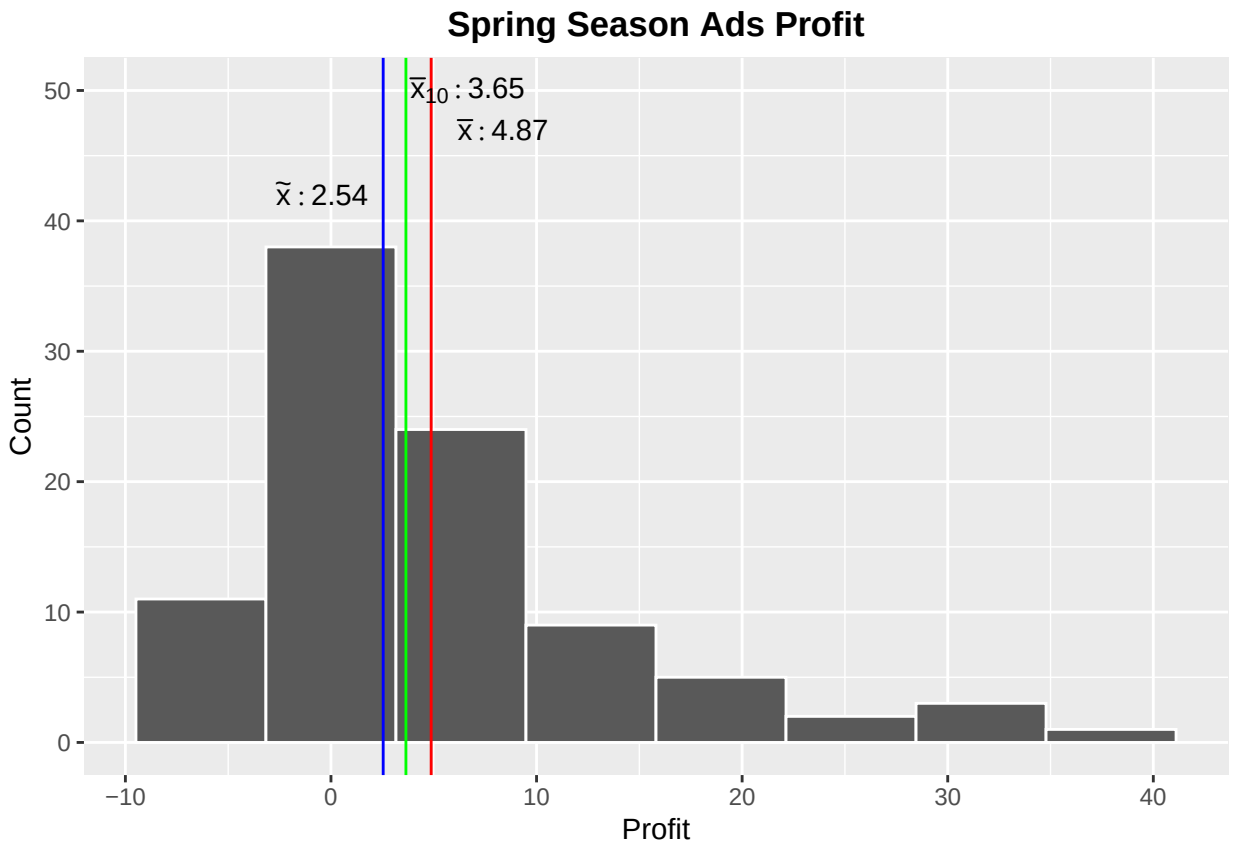
```
data = read.csv("ads4.csv")
data["Profit"] = data$adrevenue - data$adcost

# Part a)
fall_data = filter(data, season == "fall")["Profit"]
spring_data = filter(data, season == "spring")["Profit"]
summer_data = filter(data, season == "summer")["Profit"]
winter_data = filter(data, season == "winter")["Profit"]

# Fall Data
```

```r
bins <- nclass.Sturges(fall_data$Profit)
ggplot(fall_data, aes(x = Profit)) + geom_histogram(bins = bins, col = "white") -> plot1
plot1 + labs(title = "Fall Season Ads Profit", x = "Profit", y = "Count") -> plot2
plot2 + theme(plot.title = element_text(hjust = 0.5, face = "bold")) -> plot3

mean <- mean(as.numeric(unlist(fall_data)))
median <- median(as.numeric(unlist(fall_data)))
tenPerTrimmedMean <- mean(as.numeric(unlist(fall_data)), trim = 0.1)
plot3 + geom_vline(aes(xintercept = mean), col = "red") + annotate("text", x = mean + 6, y = 75,
    parse = TRUE, label = paste("bar(x) :", round(mean, 2))) + geom_vline(aes(xintercept = median),
    col = "blue") + annotate("text", x = median - 7, y = 70, parse = TRUE, label = paste("tilde(x) :",
    median)) + geom_vline(aes(xintercept = tenPerTrimmedMean), col = "green") + annotate("text",
    x = tenPerTrimmedMean + 7, y = 80, parse = TRUE, label = paste("bar(x)[10] :", round(tenPerTrimmedM
        2)))
```

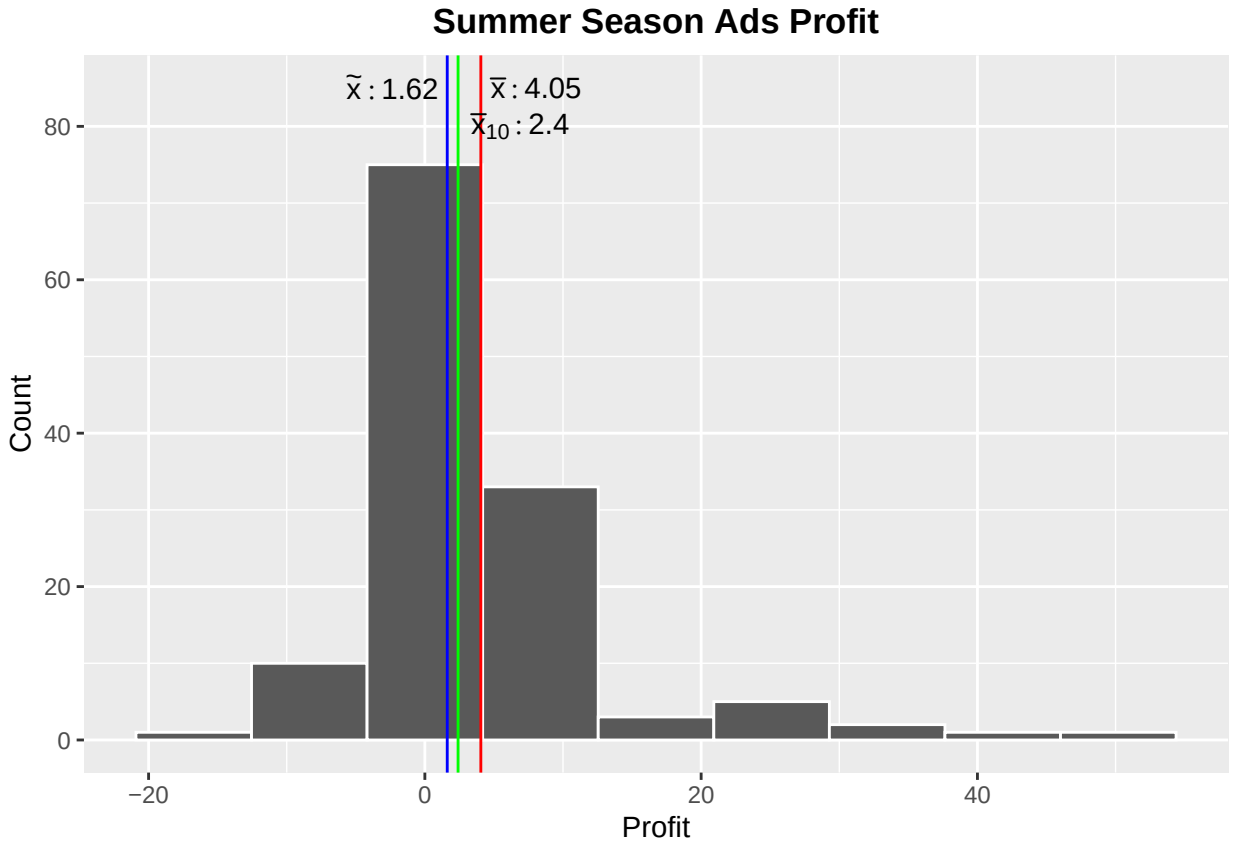



```r
# Spring Data

bins <- nclass.Sturges(spring_data$Profit)
ggplot(spring_data, aes(x = Profit)) + geom_histogram(bins = bins, col = "white") -> plot1
plot1 + labs(title = "Spring Season Ads Profit", x = "Profit", y = "Count") -> plot2
plot2 + theme(plot.title = element_text(hjust = 0.5, face = "bold")) -> plot3

mean <- mean(as.numeric(unlist(spring_data)))
median <- median(as.numeric(unlist(spring_data)))
tenPerTrimmedMean <- mean(as.numeric(unlist(spring_data)), trim = 0.1)
plot3 + geom_vline(aes(xintercept = mean), col = "red") + annotate("text", x = mean + 3.5,
    y = 47, parse = TRUE, label = paste("bar(x) :", round(mean, 2))) + geom_vline(aes(xintercept = media
```

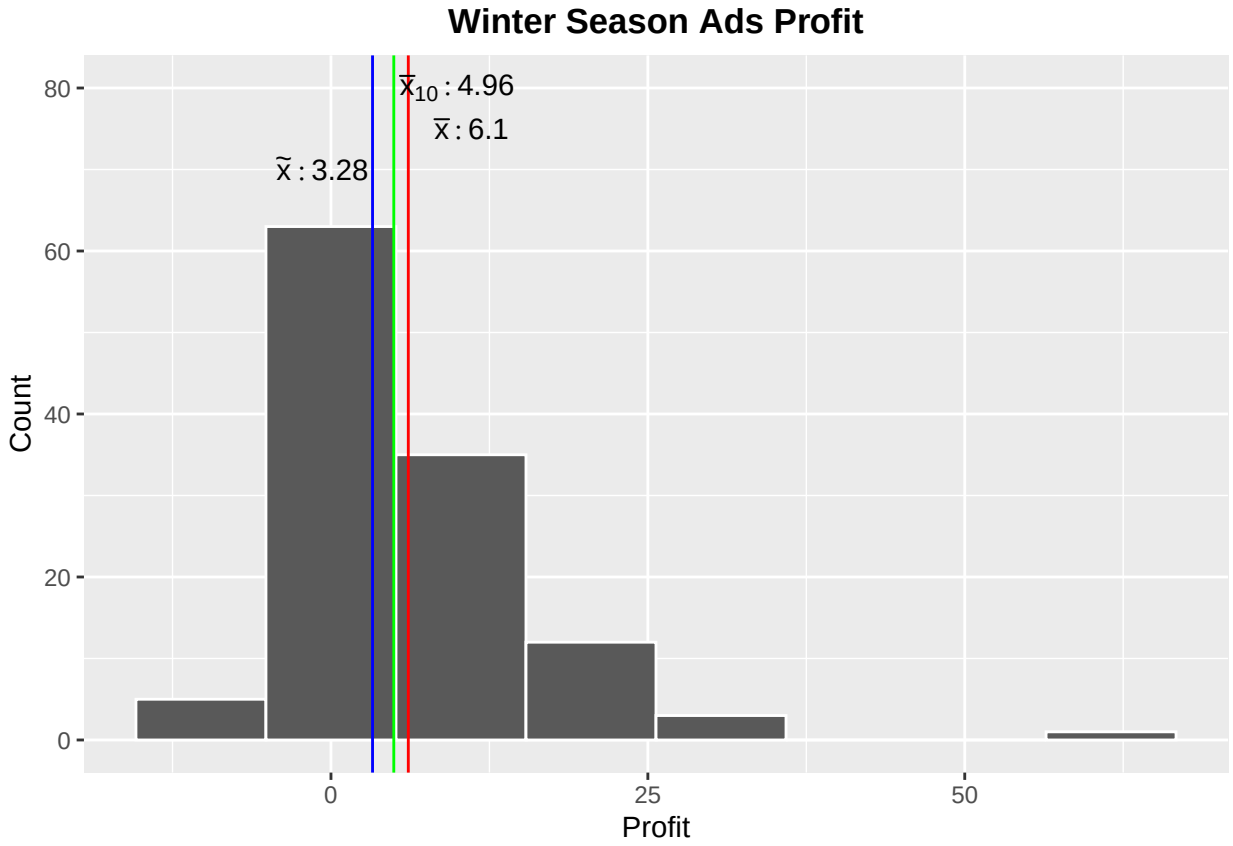
```
    col = "blue") + annotate("text", x = median - 3, y = 42, parse = TRUE, label = paste("tilde(x) :",
    median)) + geom_vline(aes(xintercept = tenPerTrimmedMean), col = "green") + annotate("text",
    x = tenPerTrimmedMean + 3, y = 50, parse = TRUE, label = paste("bar(x)[10] :", round(tenPerTrimmedM
        2)))
```




```
# Summer Data

bins <- nclass.Sturges(summer_data$Profit)
ggplot(summer_data, aes(x = Profit)) + geom_histogram(bins = bins, col = "white") -> plot1
plot1 + labs(title = "Summer Season Ads Profit", x = "Profit", y = "Count") -> plot2
plot2 + theme(plot.title = element_text(hjust = 0.5, face = "bold")) -> plot3

mean <- mean(as.numeric(unlist(summer_data)))
median <- median(as.numeric(unlist(summer_data)))
tenPerTrimmedMean <- mean(as.numeric(unlist(summer_data)), trim = 0.1)
plot3 + geom_vline(aes(xintercept = mean), col = "red") + annotate("text", x = mean + 4, y = 85,
    parse = TRUE, label = paste("bar(x) :", round(mean, 2))) + geom_vline(aes(xintercept = median),
    col = "blue") + annotate("text", x = median - 4, y = 85, parse = TRUE, label = paste("tilde(x) :",
    median)) + geom_vline(aes(xintercept = tenPerTrimmedMean), col = "green") + annotate("text",
    x = tenPerTrimmedMean + 4.5, y = 80, parse = TRUE, label = paste("bar(x)[10] :", round(tenPerTrimme
        2)))
```




```
# Winter Data

bins <- nclass.Sturges(winter_data$Profit)
ggplot(winter_data, aes(x = Profit)) + geom_histogram(bins = bins, color = "white") -> plot1
plot1 + labs(title = "Winter Season Ads Profit", x = "Profit", y = "Count") -> plot2
plot2 + theme(plot.title = element_text(hjust = 0.5, face = "bold")) -> plot3

mean <- mean(as.numeric(unlist(winter_data)))
median <- median(as.numeric(unlist(winter_data)))
tenPerTrimmedMean <- mean(as.numeric(unlist(winter_data)), trim = 0.1)
plot3 + geom_vline(aes(xintercept = mean), col = "red") + annotate("text", x = mean + 5, y = 75,
    parse = TRUE, label = paste("bar(x) :", round(mean, 2))) + geom_vline(aes(xintercept = median),
    col = "blue") + annotate("text", x = median - 4, y = 70, parse = TRUE, label = paste("tilde(x) :",
    median)) + geom_vline(aes(xintercept = tenPerTrimmedMean), col = "green") + annotate("text",
    x = tenPerTrimmedMean + 5, y = 80, parse = TRUE, label = paste("bar(x)[10] :", round(tenPerTrimmedM
        2)))
```




```
paste(" Insight : Distribution of age across all seasons is slighly right skewed with mean value greate
```

[1] " Insight : Distribution of age across all seasons is slighly right skewed with mean value greater than median. Trimmed mean gives a better picture as it elimates 10% data from both ends"

```
# Part b) H0 - summer ads yield less than or equal to profit than winter ads H1 - summer
# ads yield more profit than winter ads

summer_data = filter(data, season == "summer")
summer_data = summer_data$adrevenue - summer_data$adcost
winter_data = filter(data, season == "winter")
winter_data = winter_data$adrevenue - winter_data$adcost

mu_summer_data_profits = mean(summer_data)
mu_winter_data_profits = mean(winter_data)

var_summer_data_profits = var(summer_data)
var_winter_data_profits = var(winter_data)

# Assuming equal population variances (unknown)
t.test(summer_data, winter_data, paired = F, var.equal = T, alternative = "greater", conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  summer_data and winter_data
## t = -1.6383, df = 248, p-value = 0.9487
## alternative hypothesis: true difference in means is greater than 0
```

```
## 95 percent confidence interval:
##  -4.114643       Inf
## sample estimates:
## mean of x mean of y
##  4.050992  6.100336
```

```
# Assuming unequal population variances
t.test(summer_data, winter_data, paired = F, var.equal = F, alternative = "greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  summer_data and winter_data
## t = -1.6376, df = 245.33, p-value = 0.9486
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -4.115512       Inf
## sample estimates:
## mean of x mean of y
##  4.050992  6.100336
```

```
paste("P-value = 0.9487, thus we fail to reject null hypothesis. Insight: We don't have enough evidence
```

[1] "P-value = 0.9487, thus we fail to reject null hypothesis. Insight: We don't have enough evidence to conclude that average summer profit is greater than winter."

```
# Part c) H0 - average profit of all pair of seasons is equal H1 - atleast one pair of
# seasons has different average profit
data["Profit"] = data$adrevenue - data$adcost

model = aov(Profit ~ season, data = data)
summary(model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## season        3    331   110.2   0.897  0.443
## Residuals   442  54328   122.9
```

```
pairwise.t.test(data$Profit, data$season, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data$Profit and data$season
##
##        fall spring summer
## spring 1.00 -      -
## summer 1.00 1.00   -
## winter 1.00 1.00   0.87
##
## P value adjustment method: bonferroni
```

```
ScheffeTest(data$Profit, data$season, conf.level = 0.95)
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $g
```

```
##                     diff    lwr.ci   upr.ci   pval
## spring-fall   -1.0331162 -5.483518 3.417285 0.9351
## summer-fall   -1.8550271 -5.952206 2.242152 0.6564
## winter-fall    0.1943167 -3.992805 4.381439 0.9994
## summer-spring -0.8219109 -5.040606 3.396784 0.9602
## winter-spring  1.2274329 -3.078668 5.533534 0.8872
## winter-summer  2.0493438 -1.890619 5.989306 0.5463
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

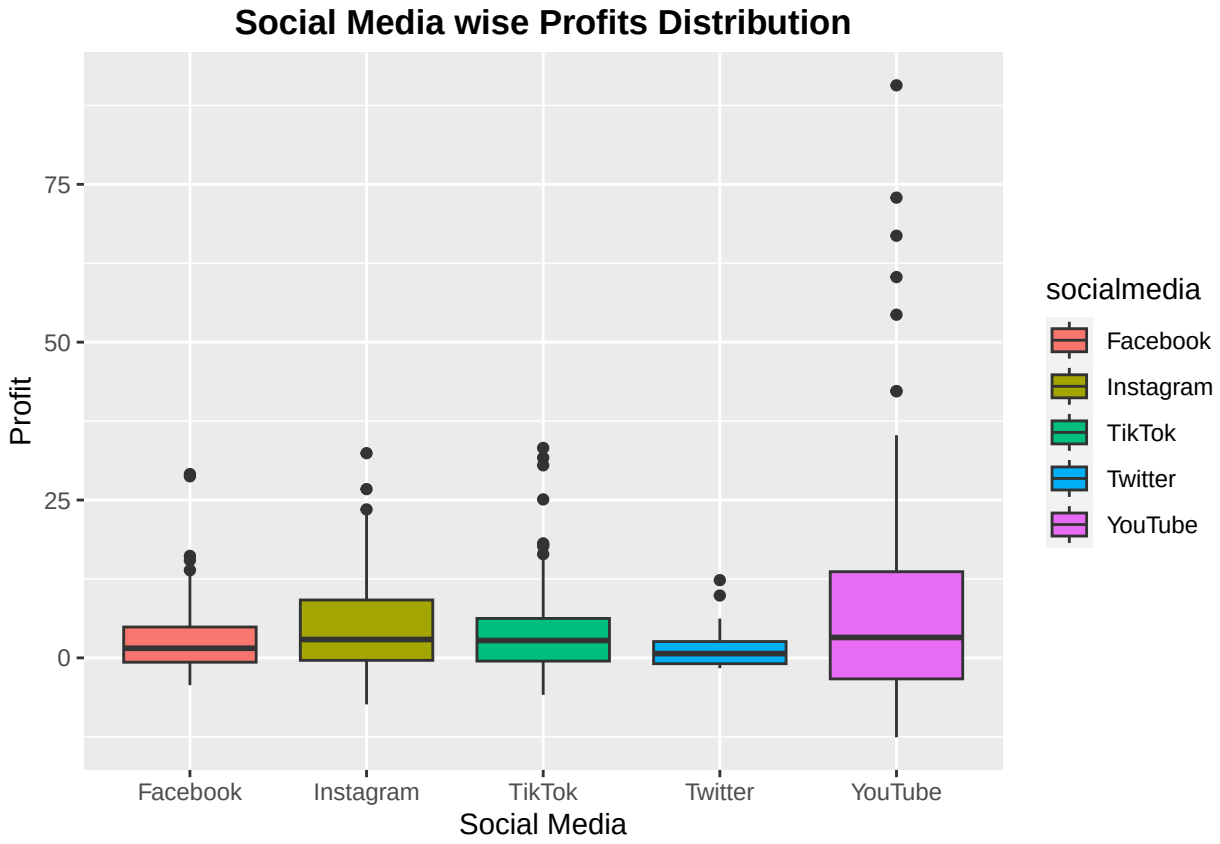
```
paste("Insight : As p value for all pair of seasons is significant (greater than 0.05), thus all pair o
```

[1] "Insight : As p value for all pair of seasons is significant (greater than 0.05), thus all pair of seasons have equal average profit at 95% significance. We conclude that choice of season does not impact profit."

Q3. The CFO wants to know if the mean profits (ad revenue - ad cost) are the same on each platform, but he adds the stipulation that the Type I error of the analysis can be at most 5%, and the familywise error of any follow-up tests can also be at most 5%. If mean profits are indeed not equal on each platform, please identify pairs of platforms for which there is a statistically significant difference between mean profits.

a. Visualize the mean profits for each platform using side-by-side boxplots. Identify any outliers and comment on trends.
b. Perform an appropriate statistical test (or series of statistical tests) and comment on your findings.

```
data = read.csv("ads4.csv")
data["Profit"] = data$adrevenue - data$adcost

# Part a)
my.bp <<- ggplot(data = data, aes(y = Profit, x = socialmedia, fill = socialmedia))
my.bp <- my.bp + geom_boxplot()
my.bp <- my.bp + ggtitle("Social Media wise Profits Distribution")
my.bp <- my.bp + ylab("Profit") + xlab("Social Media")
my.bp <- my.bp + theme(plot.title = element_text(hjust = 0.5, vjust = 0.5, face = "bold"))
my.bp
```




```
paste(" Insight : Mean profit for Instagram is more than Tiktok with less outliers so company can consi
```

[1] " Insight : Mean profit for Instagram is more than Tiktok with less outliers so company can consider investing more on ads over Instagram as compared to Tiktok. Both have approximately 30% negative profit generating ads. Twitter has the least mean profit and % ads distribution. Distribution of profit at Youtube is right skewed. Around 38% Youtube ads are generating negative profit. Overall, youtube ads have higher variance and more outliers on the positive side that is essentially leading to a higher avg profit as other platforms i.e. 8"

```
# Part b) H0 - average profit of social media platforms is same H1 - atleast one pair of
# social media platforms has different average profit
model = aov(Profit ~ socialmedia, data = data)
summary(model)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## socialmedia   4   1639   409.8   3.408 0.00925 **
## Residuals   441  53019   120.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
paste("P-val = 0.00925, thus we reject the null hypothesis. Insight: We can conclude that mean profit b
```

[1] "P-val = 0.00925, thus we reject the null hypothesis. Insight: We can conclude that mean profit between various social media platforms is different which aligns with our observations above."

Q4. The CFO also wants to better understand the relationship between acquiring new customers and net profit.

a. Visualize the relationship between whether or not someone is a new customer and the net profit off of that customer using an appropriate tool from descriptive statistics. Comment on any trends you

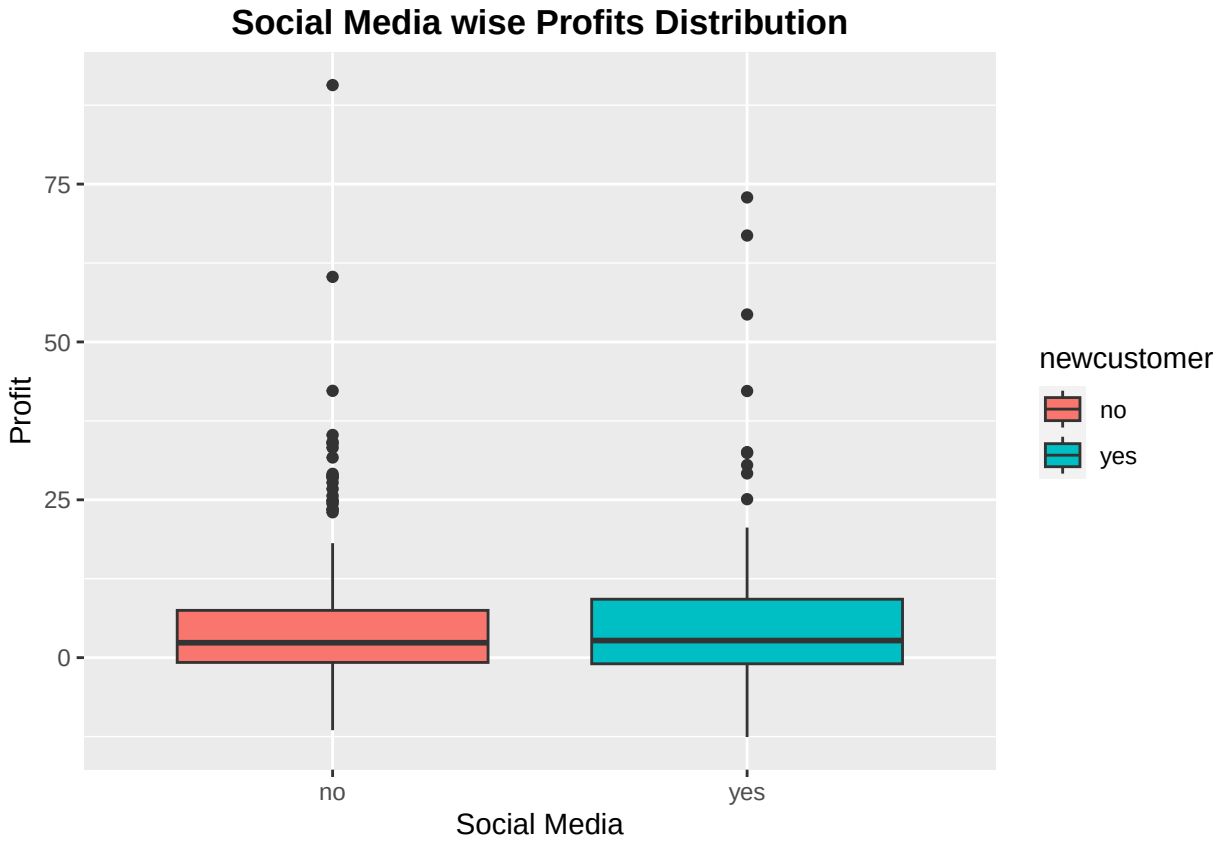
observe.

b. Is advertising on different social media platforms associated with different rates of acquiring new customers? Run an appropriate statistical test at the alpha = 0.05 significance level and comment on the results.

c. Construct a two-sided 95% confidence interval for the proportion of ads that lead to new customers.

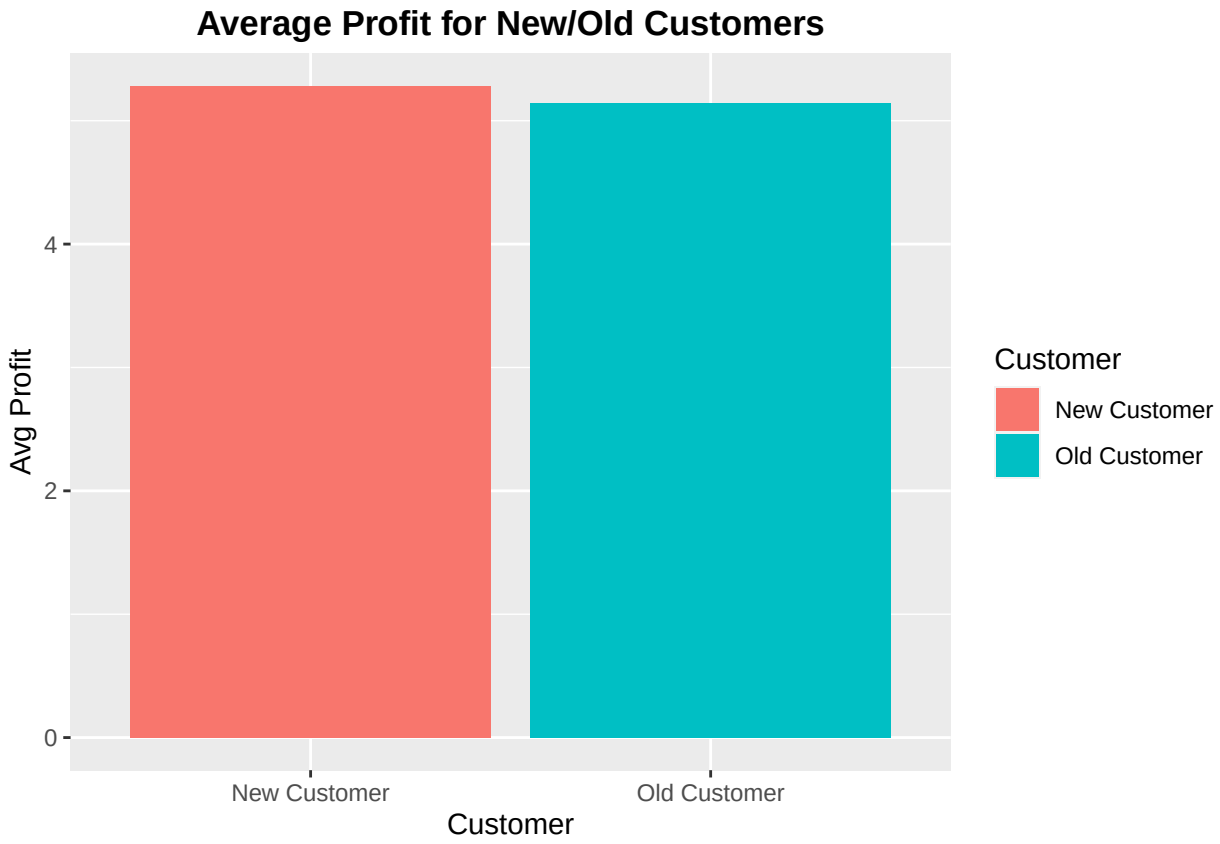
d. An analyst on another team claims that acquiring new customers is more profitable than trying to sell more products to existing customers. Test their claim at the alpha = 0.05 significance level, and comment on your results.

```
# Part a)
data = read.csv("ads4.csv")
data["Profit"] = data$adrevenue - data$adcost
new_customers = filter(data, newcustomer == "yes")["Profit"]
new_customers_avg_profit = mean(as.numeric(unlist(new_customers)))
new_customers_median_profit = median(as.numeric(unlist(new_customers)))
old_customers = filter(data, newcustomer == "no")["Profit"]
old_customers_avg_profit = mean(as.numeric(unlist(old_customers)))
old_customers_median_profit = median(as.numeric(unlist(old_customers)))

my.bp <<- ggplot(data = data, aes(y = Profit, x = newcustomer, fill = newcustomer))
my.bp <- my.bp + geom_boxplot()
my.bp <- my.bp + ggtitle("Social Media wise Profits Distribution")
my.bp <- my.bp + ylab("Profit") + xlab("Social Media")
my.bp <- my.bp + theme(plot.title = element_text(hjust = 0.5, vjust = 0.5, face = "bold"))
my.bp
```

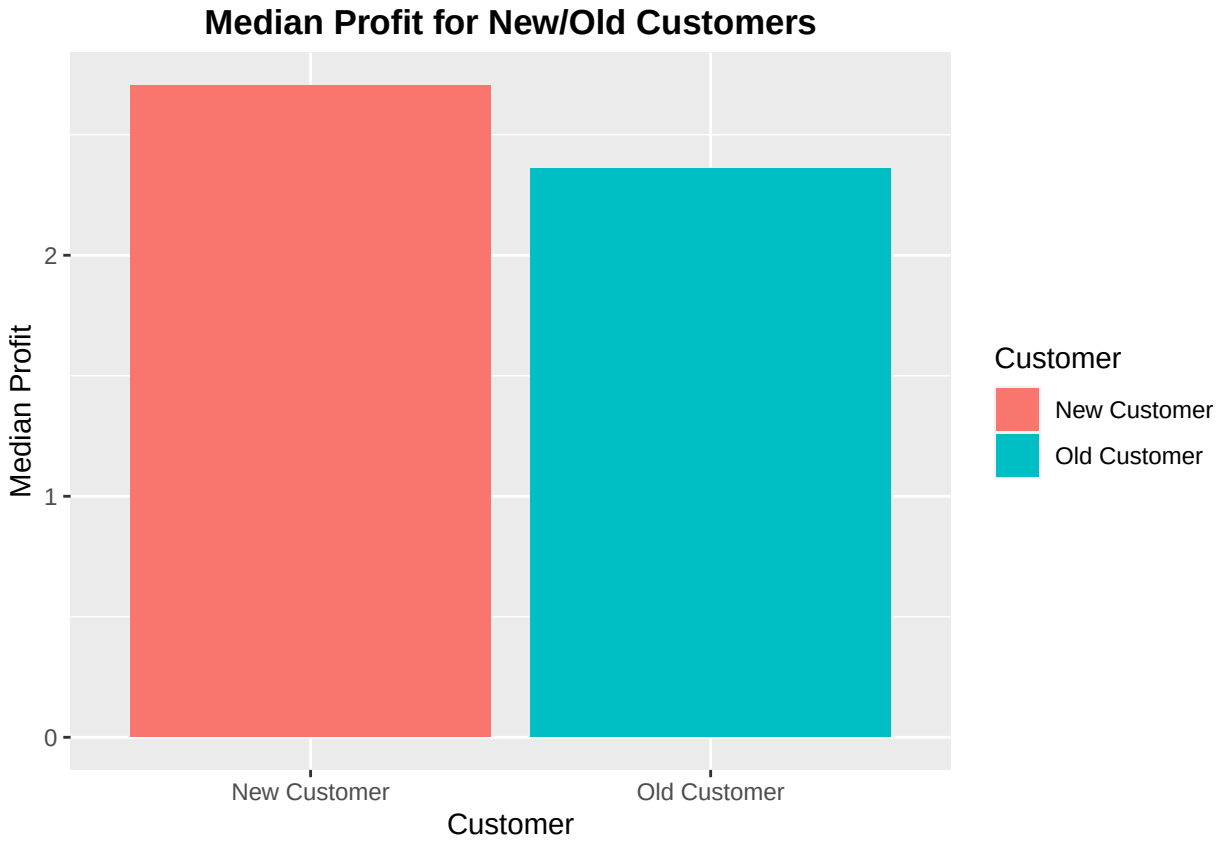



```
d = data.frame(Profit = c(new_customers_avg_profit, old_customers_avg_profit), Customer = c("New Custome
    "Old Customer"))
```

```
ggplot(d, aes(x = Customer, y = Profit, fill = Customer)) + geom_bar(stat = "identity", position = "dod
    labs(title = "Average Profit for New/Old Customers", x = "Customer", y = "Avg Profit") +
    theme(plot.title = element_text(hjust = 0.5, vjust = 0.5, face = "bold"))
```




```
d = data.frame(Profit = c(new_customers_median_profit, old_customers_median_profit), Customer = c("New
    "Old Customer"))
ggplot(d, aes(x = Customer, y = Profit, fill = Customer)) + geom_bar(stat = "identity", position = "dod
    labs(title = "Median Profit for New/Old Customers", x = "Customer", y = "Median Profit") +
    theme(plot.title = element_text(hjust = 0.5, vjust = 0.5, face = "bold"))
```




```
paste(" Insight : Overall new customers are generating more profit than old customers but since the gap
```

[1] " Insight : Overall new customers are generating more profit than old customers but since the gap is very less we will rely on significant test done at a later stage. At Youtube & Facebook, old customers are generating more profit whereas it is the opp case at Insta(new). Also, at Insta 30% are new users. Therefore, putting more ads on insta and acquiring more customers can generate more profit. For Tiktok, profit is same for both category of customers"

```
# Part b) H0 - rate of acquiring new users for social media platforms is same, i.e.
# proportion of new users are same across any two social media platforms in study, H1 -
# rate of acquiring new users for social media platforms is different, i.e., different
# proportions of new users across any two social media platforms in study.
twitter_data = filter(data, socialmedia == "Twitter")
instagram_data = filter(data, socialmedia == "Instagram")
facebook_data = filter(data, socialmedia == "Facebook")
tiktok_data = filter(data, socialmedia == "TikTok")
youtube_data = filter(data, socialmedia == "YouTube")

twitter_data_new_cust_count = nrow(filter(twitter_data, newcustomer == "yes"))
instagram_data_new_cust_count = nrow(filter(instagram_data, newcustomer == "yes"))
facebook_data_new_cust_count = nrow(filter(facebook_data, newcustomer == "yes"))
tiktok_data_new_cust_count = nrow(filter(tiktok_data, newcustomer == "yes"))
youtube_data_new_cust_count = nrow(filter(youtube_data, newcustomer == "yes"))

prop.test(x = c(twitter_data_new_cust_count, instagram_data_new_cust_count), n = c(nrow(twitter_data),
    nrow(instagram_data)), p = NULL)
```

```
##
```

```
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(twitter_data_new_cust_count, instagram_data_new_cust_count) out of c(nrow(twitter_data), nr
## X-squared = 0.083258, df = 1, p-value = 0.7729
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.3344951  0.1917500
## sample estimates:
##    prop 1    prop 2
## 0.2352941 0.3066667
```

```
prop.test(x = c(twitter_data_new_cust_count, facebook_data_new_cust_count), n = c(nrow(twitter_data),
    nrow(facebook_data)), p = NULL)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(twitter_data_new_cust_count, facebook_data_new_cust_count) out of c(nrow(twitter_data), nrow
## X-squared = 1.2603, df = 1, p-value = 0.2616
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1129814  0.3863865
## sample estimates:
##     prop 1     prop 2
## 0.23529412 0.09859155
```

```
prop.test(x = c(twitter_data_new_cust_count, tiktok_data_new_cust_count), n = c(nrow(twitter_data),
    nrow(tiktok_data)), p = NULL)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(twitter_data_new_cust_count, tiktok_data_new_cust_count) out of c(nrow(twitter_data), nrow(
## X-squared = 2.4413, df = 1, p-value = 0.1182
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.48098767  0.01824257
## sample estimates:
##    prop 1    prop 2
## 0.2352941 0.4666667
```

```
prop.test(x = c(twitter_data_new_cust_count, youtube_data_new_cust_count), n = c(nrow(twitter_data),
    nrow(youtube_data)), p = NULL)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(twitter_data_new_cust_count, youtube_data_new_cust_count) out of c(nrow(twitter_data), nrow
## X-squared = 2.0559, df = 1, p-value = 0.1516
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.46766083  0.03599343
## sample estimates:
##    prop 1    prop 2
## 0.2352941 0.4511278
```

```r
prop.test(x = c(instagram_data_new_cust_count, facebook_data_new_cust_count), n = c(nrow(instagram_data
    nrow(facebook_data)), p = NULL)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(instagram_data_new_cust_count, facebook_data_new_cust_count) out of c(nrow(instagram_data),
## X-squared = 8.4398, df = 1, p-value = 0.003671
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.06907143 0.34707880
## sample estimates:
##     prop 1     prop 2
## 0.30666667 0.09859155
```

```r
prop.test(x = c(instagram_data_new_cust_count, tiktok_data_new_cust_count), n = c(nrow(instagram_data),
    nrow(tiktok_data)), p = NULL)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(instagram_data_new_cust_count, tiktok_data_new_cust_count) out of c(nrow(instagram_data), ni
## X-squared = 4.6394, df = 1, p-value = 0.03125
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.30139395 -0.01860605
## sample estimates:
##    prop 1    prop 2
## 0.3066667 0.4666667
```

```r
prop.test(x = c(instagram_data_new_cust_count, youtube_data_new_cust_count), n = c(nrow(instagram_data)
    nrow(youtube_data)), p = NULL)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(instagram_data_new_cust_count, youtube_data_new_cust_count) out of c(nrow(instagram_data), n
## X-squared = 3.5927, df = 1, p-value = 0.05803
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2892084510  0.0002861452
## sample estimates:
##    prop 1    prop 2
## 0.3066667 0.4511278
```

```r
prop.test(x = c(facebook_data_new_cust_count, tiktok_data_new_cust_count), n = c(nrow(facebook_data),
    nrow(tiktok_data)), p = NULL)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(facebook_data_new_cust_count, tiktok_data_new_cust_count) out of c(nrow(facebook_data), nrow
## X-squared = 27.16, df = 1, p-value = 1.873e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.4841974 -0.2519529
```

```
## sample estimates:
##     prop 1     prop 2
## 0.09859155 0.46666667
```

```
prop.test(x = c(facebook_data_new_cust_count, youtube_data_new_cust_count), n = c(nrow(facebook_data),
    nrow(youtube_data)), p = NULL)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(facebook_data_new_cust_count, youtube_data_new_cust_count) out of c(nrow(facebook_data), nr
## X-squared = 24.509, df = 1, p-value = 7.397e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.4727005 -0.2323720
## sample estimates:
##     prop 1     prop 2
## 0.09859155 0.45112782
```

```
prop.test(x = c(tiktok_data_new_cust_count, youtube_data_new_cust_count), n = c(nrow(tiktok_data),
    nrow(youtube_data)), p = NULL)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(tiktok_data_new_cust_count, youtube_data_new_cust_count) out of c(nrow(tiktok_data), nrow(y
## X-squared = 0.020249, df = 1, p-value = 0.8868
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1078543  0.1389319
## sample estimates:
##    prop 1    prop 2
## 0.4666667 0.4511278
```

```
paste(" Insight : Some of the pair of social media platforms does have equal rate of acquiring new users
```

```
[1] " Insight : Some of the pair of social media platforms does have equal rate of acquiring new users while others have different rate as can be seen from the pvalues. Therefore, we conclude advertising on different social media platforms is associated with different rates of acquiring new customers but it depends on social media paltform used. Different proportion values are seen among pairs (Instagram, Facebook), (Instagram, TikTok), (Facebook, TikTok), (Facebook, YouTube), while other pairs have significantly equal proportion of new users."
```

```
# Part c)
prop.test(x = twitter_data_new_cust_count, n = nrow(twitter_data), p = NULL, conf.level = 0.95,
    correct = TRUE)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  twitter_data_new_cust_count out of nrow(twitter_data), null probability 0.5
## X-squared = 3.7647, df = 1, p-value = 0.05235
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.07823122 0.50238373
## sample estimates:
##         p
```

```
## 0.2352941
```

```
prop.test(x = facebook_data_new_cust_count, n = nrow(facebook_data), p = NULL, conf.level = 0.95,
    correct = TRUE)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  facebook_data_new_cust_count out of nrow(facebook_data), null probability 0.5
## X-squared = 44.169, df = 1, p-value = 3.012e-11
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.04389665 0.19839452
## sample estimates:
##          p
## 0.09859155
```

```
prop.test(x = tiktok_data_new_cust_count, n = nrow(tiktok_data), p = NULL, conf.level = 0.95,
    correct = TRUE)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  tiktok_data_new_cust_count out of nrow(tiktok_data), null probability 0.5
## X-squared = 0.54, df = 1, p-value = 0.4624
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3854464 0.5496202
## sample estimates:
##         p
## 0.4666667
```

```
prop.test(x = youtube_data_new_cust_count, n = nrow(youtube_data), p = NULL, conf.level = 0.95,
    correct = TRUE)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  youtube_data_new_cust_count out of nrow(youtube_data), null probability 0.5
## X-squared = 1.0827, df = 1, p-value = 0.2981
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3655242 0.5395956
## sample estimates:
##         p
## 0.4511278
```

```
prop.test(x = instagram_data_new_cust_count, n = nrow(instagram_data), p = NULL, conf.level = 0.95,
    correct = TRUE)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  instagram_data_new_cust_count out of nrow(instagram_data), null probability 0.5
## X-squared = 10.453, df = 1, p-value = 0.001224
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
```

```
##  0.2081231 0.4252191
## sample estimates:
##         p
## 0.3066667
```

```
paste("Confidence Intervals thus found are as follows: Twitter=[0.07823122, 0.50238373], Facebook=[0.043
```

[1] "Confidence Intervals thus found are as follows: Twitter=[0.07823122, 0.50238373], Facebook=[0.04389665, 0.19839452]. TikTok=[0.3854464, 0.5496202], YouTube=[0.3655242, 0.5395956], Instagram=[0.2081231, 0.4252191]."

```
# Part d) H0 - average profit for new customers is less than or equal to that of existing
# customers H1 - average profit for new customers is more than that of existing customers
# Assuming equal population variances (unknown)
t.test(data[data$newcustomer == "yes", ]$Profit, data[data$newcustomer == "no", ]$Profit, paired = F,
    var.equal = T, alternative = "greater", conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  data[data$newcustomer == "yes", ]$Profit and data[data$newcustomer == "no", ]$Profit
## t = 0.12628, df = 444, p-value = 0.4498
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.65832      Inf
## sample estimates:
## mean of x mean of y
##  5.284573  5.146986
```

```
# Assuming unequal population variances
t.test(data[data$newcustomer == "yes", ]$Profit, data[data$newcustomer == "no", ]$Profit, paired = F,
    var.equal = F, alternative = "greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  data[data$newcustomer == "yes", ]$Profit and data[data$newcustomer == "no", ]$Profit
## t = 0.12269, df = 311.56, p-value = 0.4512
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.712467       Inf
## sample estimates:
## mean of x mean of y
##  5.284573  5.146986
```

```
paste("P-value = 0.4498, thus we fail to reject the null hypothesis, i.e., we don't have enough evidence
```

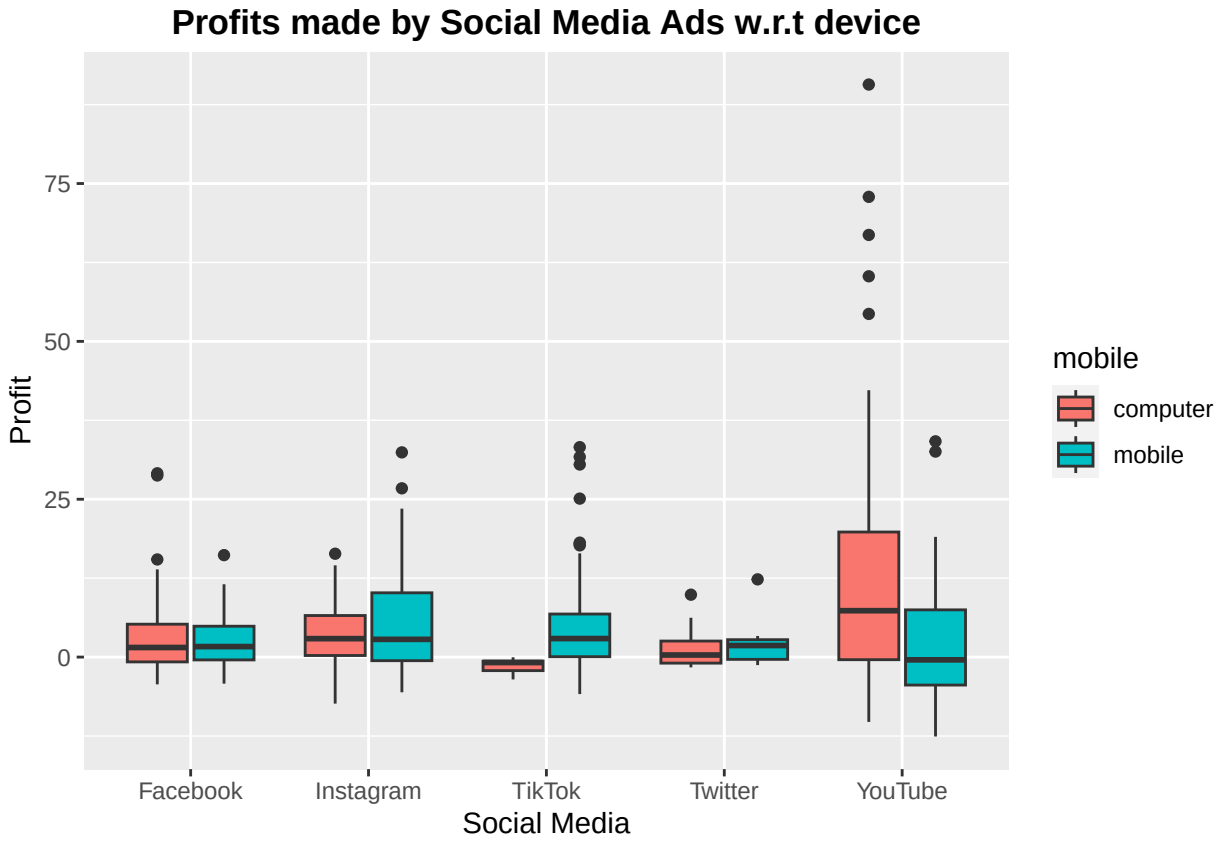
[1] "P-value = 0.4498, thus we fail to reject the null hypothesis, i.e., we don't have enough evidence to conclude that the average profit for new customers is more than that generated by the existing customers. Insight: Acquiring new customers is less profitable than selling to existing customers."

Q5. The CEO and CFO disagree about whether being on a mobile phone affects average profits. The CEO thinks that being on your phone or computer doesn't affect overall profits, whereas the CFO thinks there is a difference.

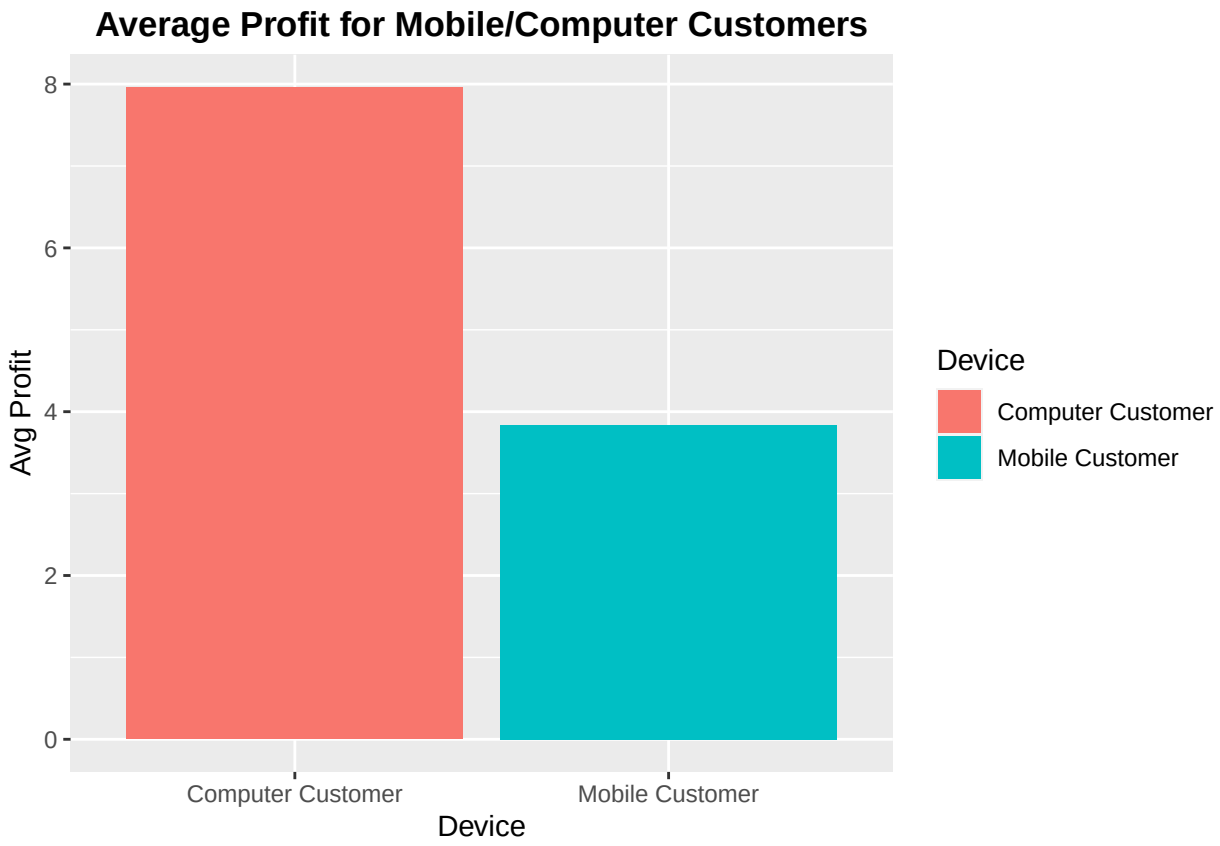
a. Visualize profits by mobile phone status for each social network platform using an appropriate tool from descriptive statistics. Make sure to label the plot (title, axes, legend), and comment on trends you observe.

b. At the alpha = 0.05 significance level, examine whether or not being on a mobile phone affects average profits for each social network platform. Discuss your findings.

```
data = read.csv("ads4.csv")
data["Profit"] = data$adrevenue - data$adcost

# Part a)
my.bp <<- ggplot(data = data, aes(y = Profit, x = socialmedia, fill = mobile))
my.bp <- my.bp + geom_boxplot()
my.bp <- my.bp + ggtitle("Profits made by Social Media Ads w.r.t device")
my.bp <- my.bp + ylab("Profit") + xlab("Social Media")
my.bp <- my.bp + theme(plot.title = element_text(hjust = 0.5, vjust = 0.5, face = "bold"))
my.bp
```




```
mobile_customers = filter(data, mobile == "mobile")["Profit"]
mobile_customers_avg_profit = mean(as.numeric(unlist(mobile_customers)))
computer_customers = filter(data, mobile == "computer")["Profit"]
computer_customers_avg_profit = mean(as.numeric(unlist(computer_customers)))

d = data.frame(Profit = c(mobile_customers_avg_profit, computer_customers_avg_profit), Device = c("Mobi
    "Computer Customer"))
ggplot(d, aes(x = Device, y = Profit, fill = Device)) + geom_bar(stat = "identity", position = "dodge")
    labs(title = "Average Profit for Mobile/Computer Customers", x = "Device", y = "Avg Profit") +
    theme(plot.title = element_text(hjust = 0.5, vjust = 0.5, face = "bold"))
```




```
paste("Insight: Avg profit for computer is greater than avg profit for mobile which is mostly driven by
```

[1] "Insight: Avg profit for computer is greater than avg profit for mobile which is mostly driven by youtube as 55% users (73) used computer driving max profit of 12.6. For tiktok, device of users affects average profits. Overall computer users are driving better average profit than mobile users"

```
# Part b) H0 - average profit for mobile is same as for computer for a social network
# platform H1 - average profit for mobile is different than that of computer for a social
# network platform
twitter_data = filter(data, socialmedia == "Twitter")
instagram_data = filter(data, socialmedia == "Instagram")
facebook_data = filter(data, socialmedia == "Facebook")
tiktok_data = filter(data, socialmedia == "TikTok")
youtube_data = filter(data, socialmedia == "YouTube")

t.test(twitter_data[twitter_data$mobile == "mobile", ]$Profit, twitter_data[twitter_data$mobile ==
    "computer", ]$Profit, paired = F, var.equal = F, alternative = "greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  twitter_data[twitter_data$mobile == "mobile", ]$Profit and twitter_data[twitter_data$mobile ==
## t = 0.38519, df = 11.19, p-value = 0.3537
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -2.969499       Inf
## sample estimates:
## mean of x mean of y
```

```
##  2.521429  1.709000
```

```
t.test(instagram_data[instagram_data$mobile == "mobile", ]$Profit, instagram_data[instagram_data$mobile
    "computer", ]$Profit, paired = F, var.equal = F, alternative = "greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  instagram_data[instagram_data$mobile == "mobile", ]$Profit and instagram_data[instagram_data$m
## t = 0.64391, df = 57.775, p-value = 0.2611
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.757404       Inf
## sample estimates:
## mean of x mean of y
##  5.411923  4.310870
```

```
t.test(facebook_data[facebook_data$mobile == "mobile", ]$Profit, facebook_data[facebook_data$mobile ==
    "computer", ]$Profit, paired = F, var.equal = F, alternative = "greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  facebook_data[facebook_data$mobile == "mobile", ]$Profit and facebook_data[facebook_data$mobil
## t = -0.95795, df = 44.314, p-value = 0.8284
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -4.267707       Inf
## sample estimates:
## mean of x mean of y
##  2.600513  4.150313
```

```
t.test(tiktok_data[tiktok_data$mobile == "mobile", ]$Profit, tiktok_data[tiktok_data$mobile ==
    "computer", ]$Profit, paired = F, var.equal = F, alternative = "greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  tiktok_data[tiktok_data$mobile == "mobile", ]$Profit and tiktok_data[tiktok_data$mobile == "co
## t = 7.8549, df = 42.788, p-value = 3.938e-10
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  4.565072      Inf
## sample estimates:
## mean of x mean of y
##  4.370775 -1.437500
```

```
t.test(youtube_data[youtube_data$mobile == "mobile", ]$Profit, youtube_data[youtube_data$mobile ==
    "computer", ]$Profit, paired = F, var.equal = F, alternative = "greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  youtube_data[youtube_data$mobile == "mobile", ]$Profit and youtube_data[youtube_data$mobile ==
## t = -3.9378, df = 110.88, p-value = 0.9999
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
```

```
##  -14.85789       Inf
## sample estimates:
## mean of x mean of y
##  2.148983 12.603243
```

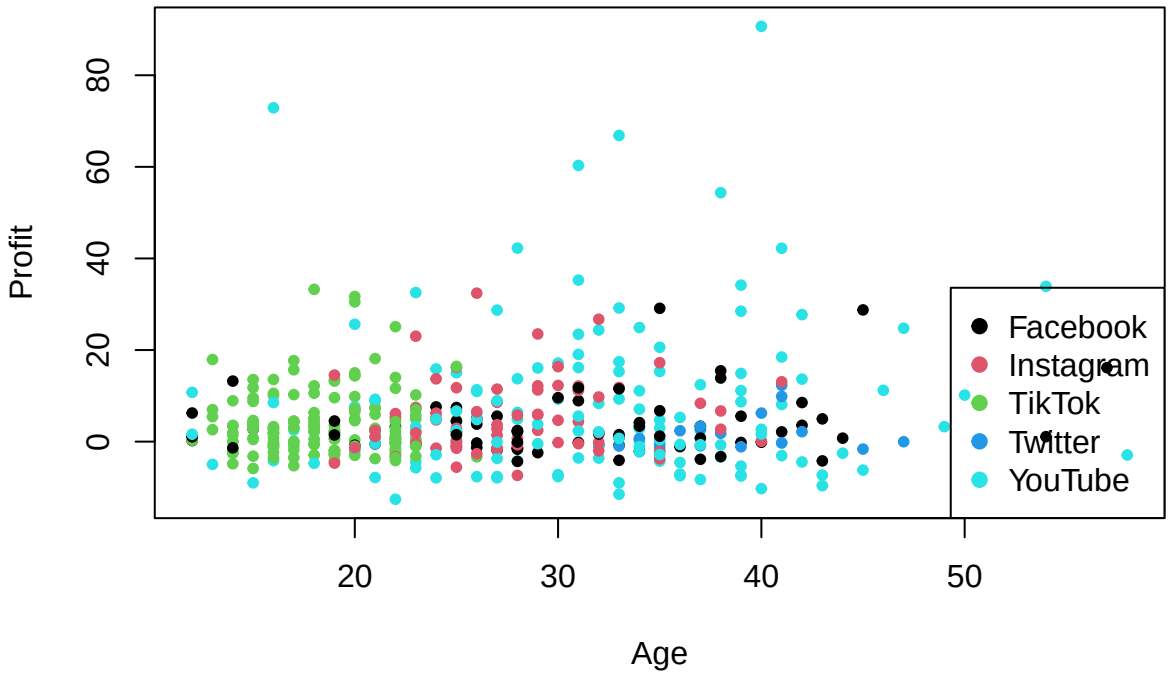
```
paste("Insight: As can be seen from the observed results, being on a mobile phone/computer does not affe
```

[1] "Insight: As can be seen from the observed results, being on a mobile phone/computer does not affect average profits for social network platform except Tiktok where p value is less than 0.05."

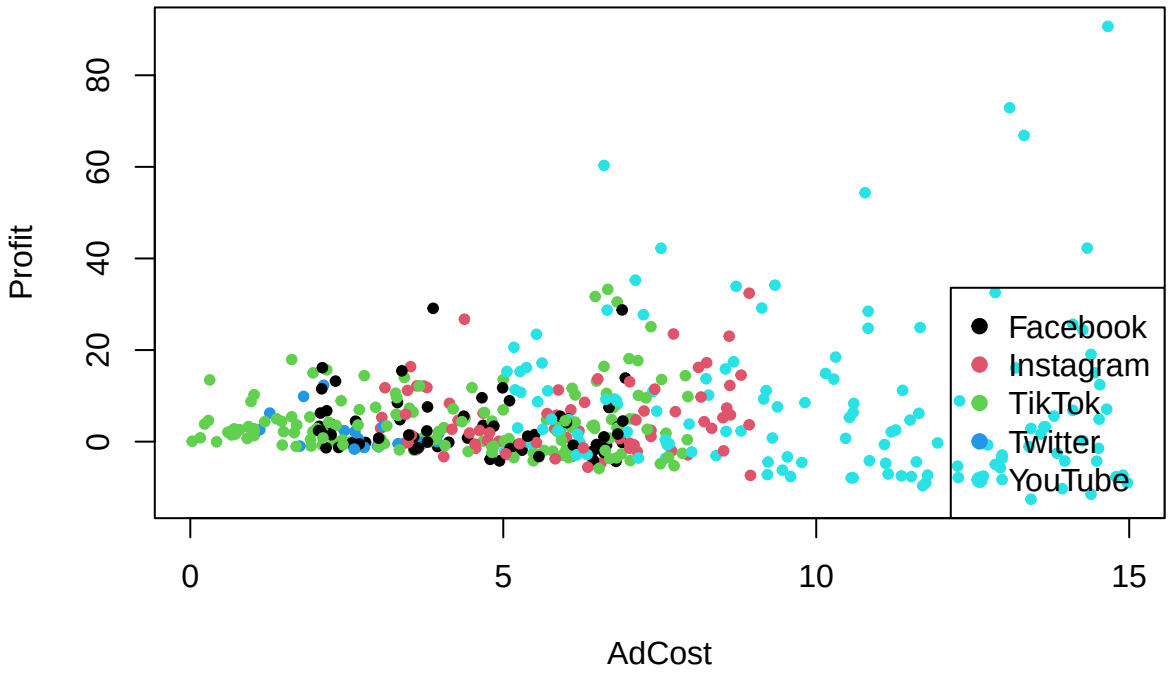
Q6. It's time to start understanding how profit depends on other variables, notably ad cost and age.

a. Visualize the relationship between profit and age with different colored points for each social network. Make sure to label the plot (title, axes, legend), and comment on trends you observe.
b. Visualize the relationship between profit and ad cost, again with different colored points for each social network. Make sure to label the plot (title, axes, legend), and comment on trends you observe.
c. Are profit and age correlated? Perform an appropriate statistical test using Pearson correlation at the alpha = 0.05 significance level and comment on the results.
d. Are profit and ad cost correlated? Perform an appropriate statistical test using Spearman correlation at the alpha = 0.05 significance level and comment on the results.
e. Fit a linear regression to model the profit as a function of ad cost. Report the regression equation, a 90% confidence interval for the coefficient of ad cost, and the coefficient of determination.
f. Fit a linear regression to model the profit as a function of ad cost and age. Comment on the results.
g. At the alpha = 0.05 significance level, conduct an F-test to determine whether ad cost significantly predicts profit once we have accounted for age. R

```
data = read.csv("ads4.csv")
data["profit"] = data$adrevenue - data$adcost

# Part a)
plot(data$age, data$profit, pch = 20, col = factor(data$socialmedia), main = "Relationship between profi
    xlab = "Age", ylab = "Profit")
# Legend Legend
legend("bottomright", legend = levels(factor(data$socialmedia)), pch = 19, col = factor(levels(factor(da
```




```
# Insights for: a.Almost all TikTok users are younger than 25 years of age, with many
# being less than 20 years.  b. YouTube has a wide age group of content creators, with
# many being the most profitable in the entire data set too. c. YouTube content creators
# seemed to also suffer the most loss across all other categories and age groups.


# Part b)
plot(data$adcost, data$profit, pch = 20, col = factor(data$socialmedia), main = "Relationship between pr
    xlab = "AdCost", ylab = "Profit")
# Legend Legend
legend("bottomright", legend = levels(factor(data$socialmedia)), pch = 19, col = factor(levels(factor(da
```



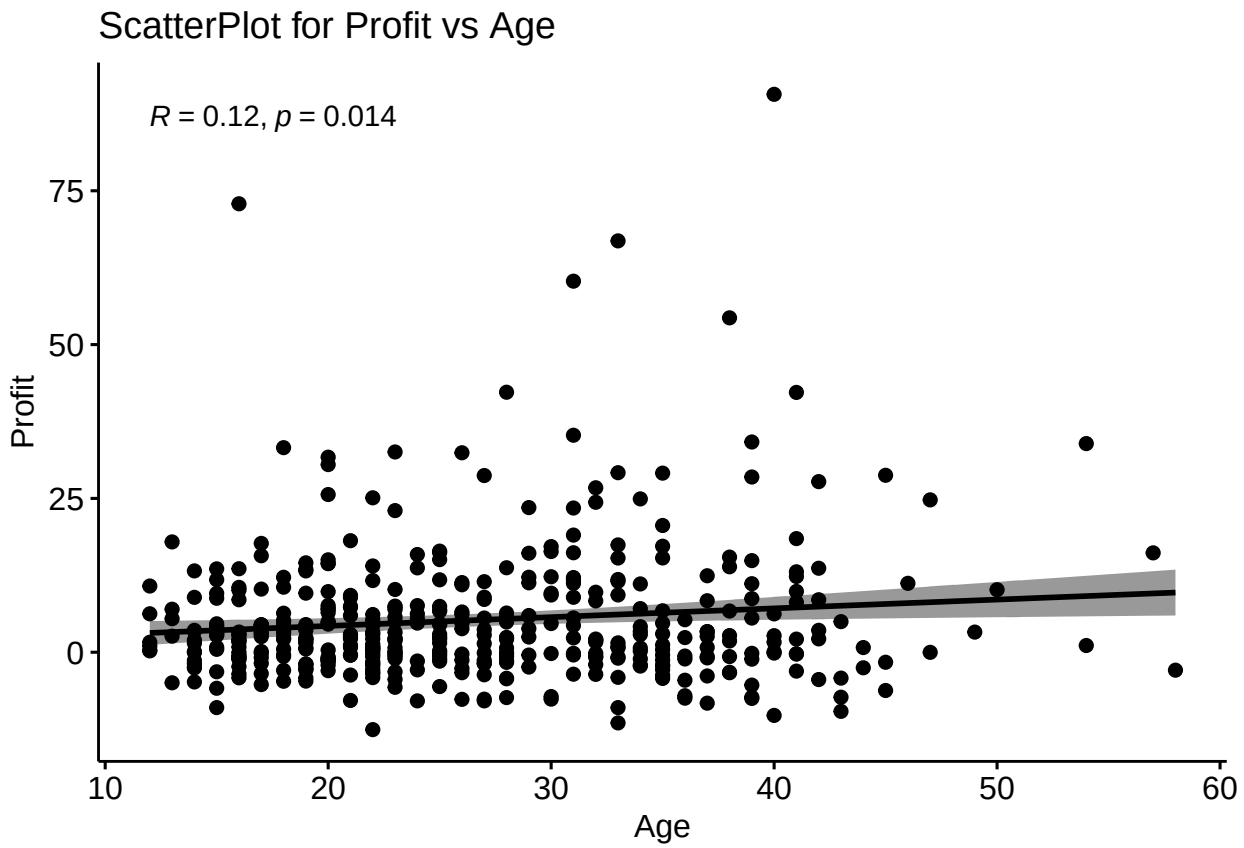

```
# Insights: a. YouTube has more losses despite having high Ad Costs. b. For other
# platforms, the AdCost is mostly limited between 0 to 10 at max. c. For other platforms,
# the profit is mostly limited between 0 to 20 at max. d. All Platforms have suffered
# with loss(<0 profit)
```

```
# Part c) H0 - profit and age are correlated H1 - profit and age are not correlated
cor.test(data$age, data$profit, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$age and data$profit
## t = 2.4655, df = 444, p-value = 0.01406
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0236149 0.2068332
## sample estimates:
##       cor
## 0.1162125
```

```
ggscatter(data, x = "age", y = "profit", add = "reg.line", conf.int = TRUE, cor.coef = TRUE,
    cor.method = "pearson", xlab = "Age", ylab = "Profit", title = "ScatterPlot for Profit vs Age")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
paste("As Pearson correlation coefficient between profit and age is just 0.1162 and p-value = 0.01406, A
```

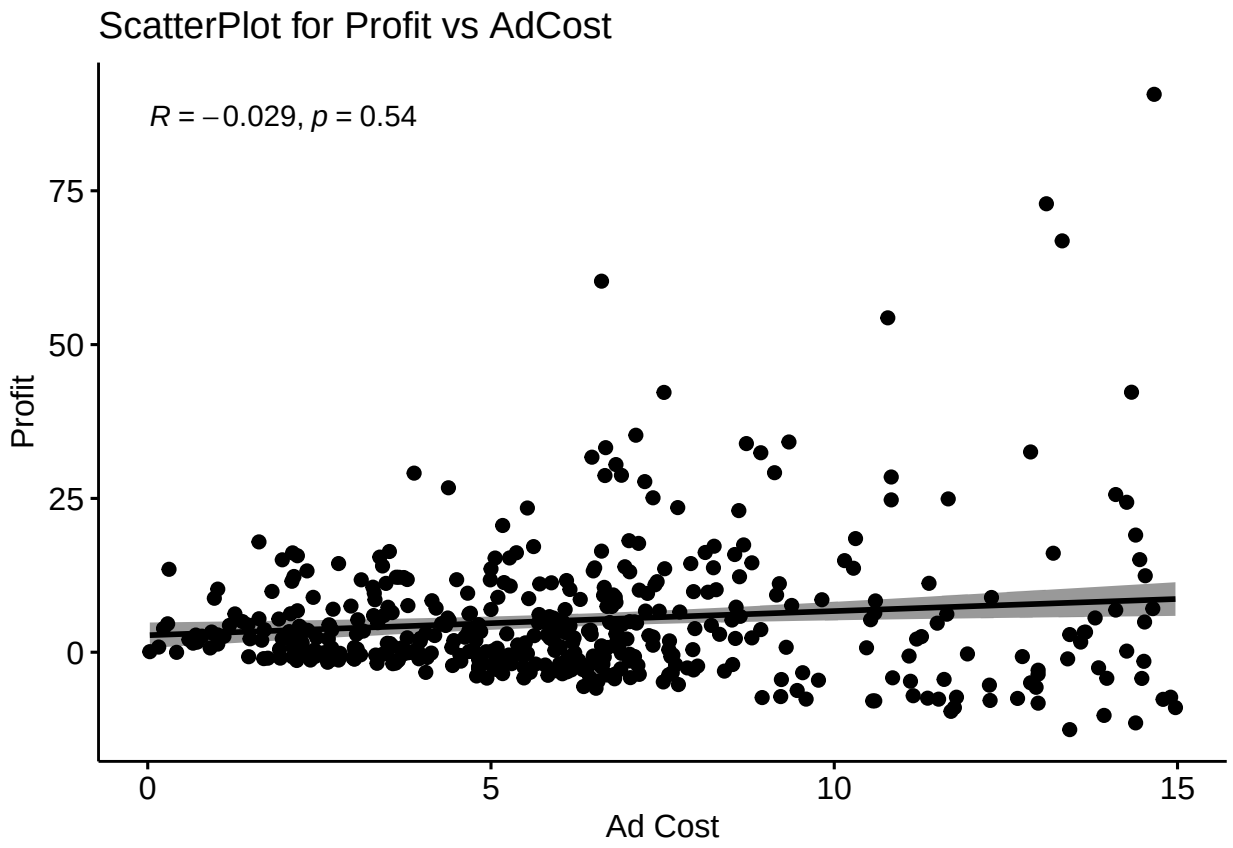
[1] "As Pearson correlation coefficient between profit and age is just 0.1162 and p-value = 0.01406, Age and Profit aren't that significantly correlated."

```
# Part d) H0 - profit and adcost are correlated H1 - profit and adcost are not correlated
cor.test(data$adcost, data$profit, method = c("spearman"))
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$adcost and data$profit
## S = 15211032, p-value = 0.5449
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.02874451
```

```
ggscatter(data, x = "adcost", y = "profit", add = "reg.line", conf.int = TRUE, cor.coef = TRUE,
    cor.method = "spearman", xlab = "Ad Cost", ylab = "Profit", title = "ScatterPlot for Profit vs AdCos
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
paste("As Spearman correlation coefficient between profit and adcost is just -0.02874 and p-value = 0.5
```

[1] "As Spearman correlation coefficient between profit and adcost is just -0.02874 and p-value = 0.5449 which is greater than 0.05, thus we don't have significant evidence to conclude that profit and adcost are not colrrelated."

```
# Part e)
model = lm(profit ~ adcost, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = profit ~ adcost, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.616  -5.642  -2.026   3.293  82.161
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7592     1.0519   2.623  0.00902 **
## adcost        0.3929     0.1472   2.669  0.00789 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.01 on 444 degrees of freedom
## Multiple R-squared:  0.01579,    Adjusted R-squared:  0.01357
## F-statistic: 7.122 on 1 and 444 DF,  p-value: 0.007895
```

```r
paste("Regression Equation: Profit = 2.75 + 0.3929 * AdCost")
```

```
## [1] "Regression Equation: Profit = 2.75 + 0.3929 * AdCost"
```

```r
print("Confidence interval is:")
```

```
## [1] "Confidence interval is:"
```

```r
confint(model, "adcost", level = 0.9)
```

```
##              5 %      95 %
## adcost 0.1502215 0.6355658
```

```r
paste("Coefficient of determination:", summary(model)$r.squared)
```

```
## [1] "Coefficient of determination: 0.0157864780194078"
```

```r
# Part f)
multi_model = lm(profit ~ adcost + age, data = data)
summary(multi_model)
```

```
##
## Call:
## lm(formula = profit ~ adcost + age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.935  -6.077  -2.018   3.207  81.408
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.45097    1.68225   0.268   0.7888
## adcost       0.31223    0.15390   2.029   0.0431 *
## age          0.10610    0.06043   1.756   0.0798 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.98 on 443 degrees of freedom
## Multiple R-squared:  0.02259,    Adjusted R-squared:  0.01817
## F-statistic: 5.119 on 2 and 443 DF,  p-value: 0.006343
```

```r
paste("Regression Equation: Profit = 0.4509 + 0.3122 * AdCost + 0.1061 * age")
```

```
## [1] "Regression Equation: Profit = 0.4509 + 0.3122 * AdCost + 0.1061 * age"
```

```r
paste("Coefficient of determination:", summary(multi_model)$r.squared)
```

```
## [1] "Coefficient of determination: 0.0225868090115788"
```

```r
# Part g) H0 - model with no independent variables fits the data as well as your model H1
# - model fits the data better than the intercept-only model
anova(model, multi_model)
```

```
## Analysis of Variance Table
##
## Model 1: profit ~ adcost
## Model 2: profit ~ adcost + age
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    444 53795
```

```
## 2    443 53424  1    371.69 3.0822 0.07985 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print("Test statistic: 3.0822 ")
```

```
## [1] "Test statistic: 3.0822 "
```

```
print("P-value: 0.07985 ")
```

```
## [1] "P-value: 0.07985 "
```

```
paste("Since the test statistic i.e. F  value is not significant with p-val as 0.07985, thus adding age
```

[1] "Since the test statistic i.e. F value is not significant with p-val as 0.07985, thus adding age to the model didn't perform better over the model with just adcost."

```
paste("In context to the problem, we can intrepret that adcost is a better indicator of profit than age
```

[1] "In context to the problem, we can intrepret that adcost is a better indicator of profit than age."

Question 7. Suppose you are given $100 to spend on advertising for this company. How would you spend it? Explain and interpret any additional analyses you want to do, and provide a detailed description of why you used the analyses you did. This portion should involve significant thought, perhaps partially based on the types of analysis you did earlier. (Hint: It may help to consider each social media platform separately.)

2022 Dec 15

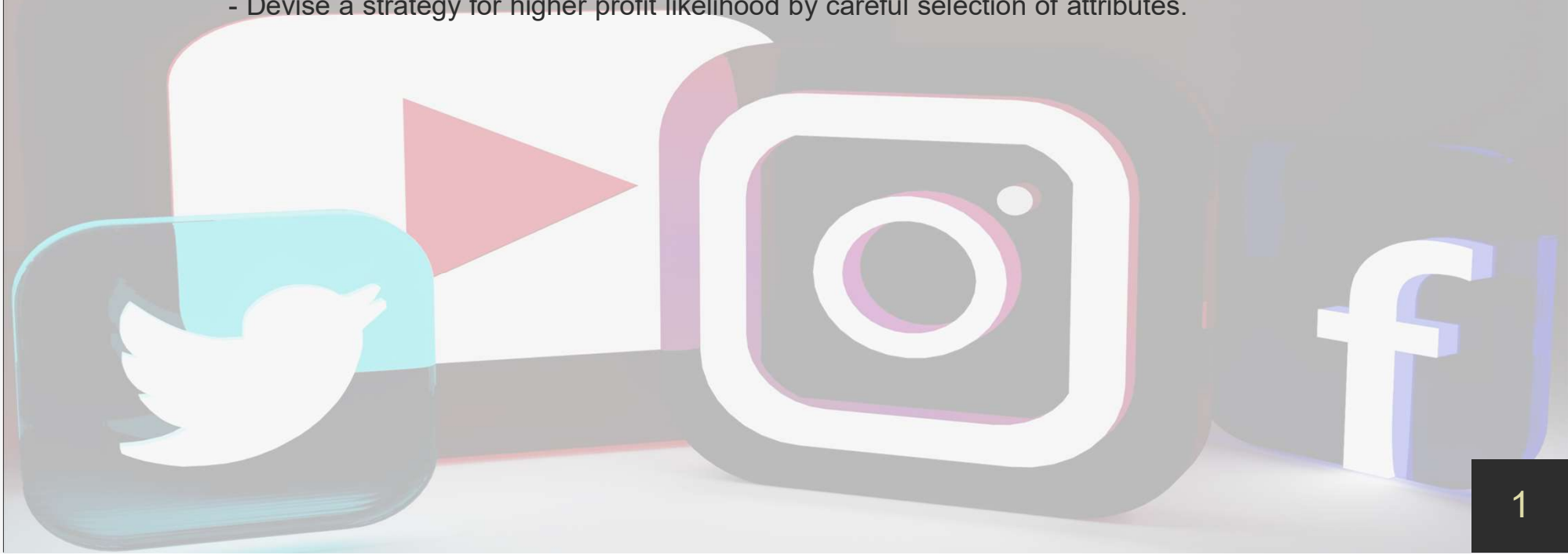
DSCC 462 Final Project

Presented by:
Team 4 - Rishabh Kandoi, Ayush Singla, Aradhya Mathur, Richa Yadav

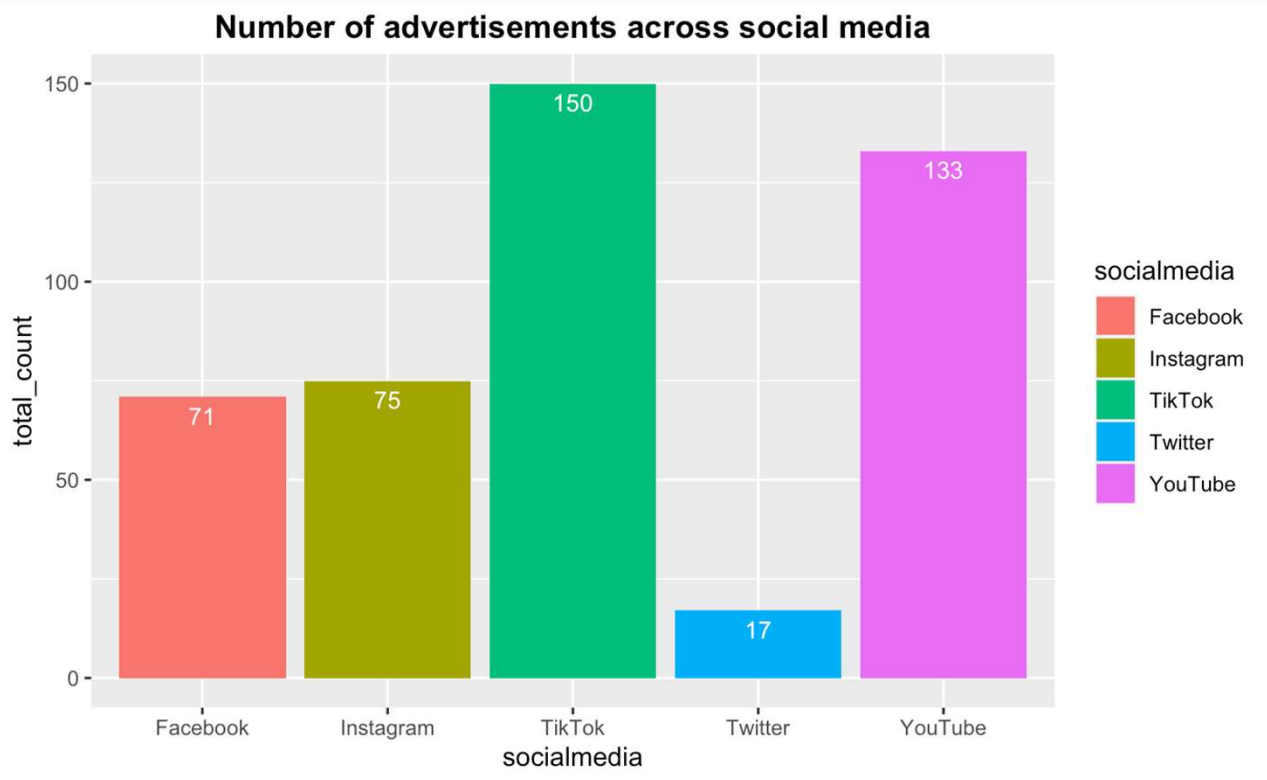
Problem Statement

Given historical data,

- Perform EDA to understand the relationship among variables and their impact on profit.
- Devise a strategy for higher profit likelihood by careful selection of attributes.

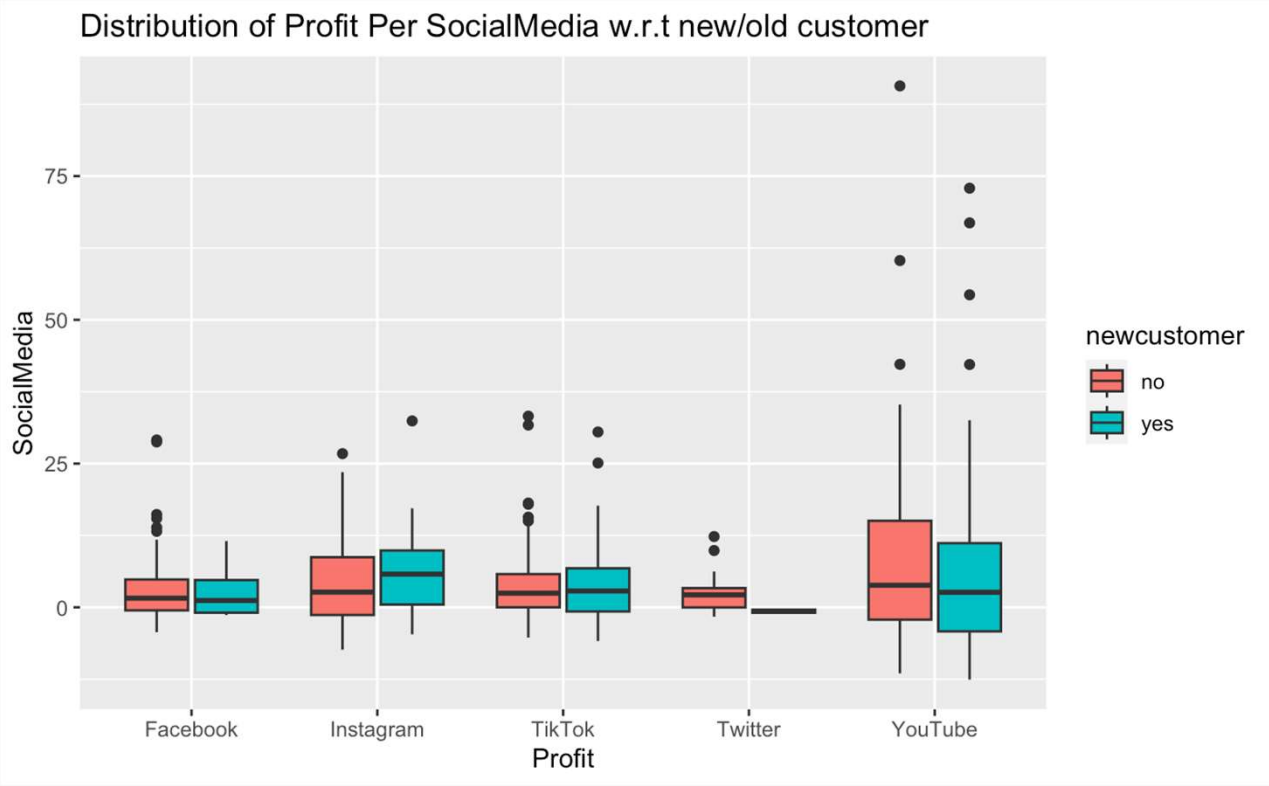


Comparison of platforms



- TikTok and YouTube together account for 60% of the total Ads running across all platforms.
- While Twitter accounts for just 4% of the entire distribution running across all platforms.

Profit across platforms w.r.t customer type



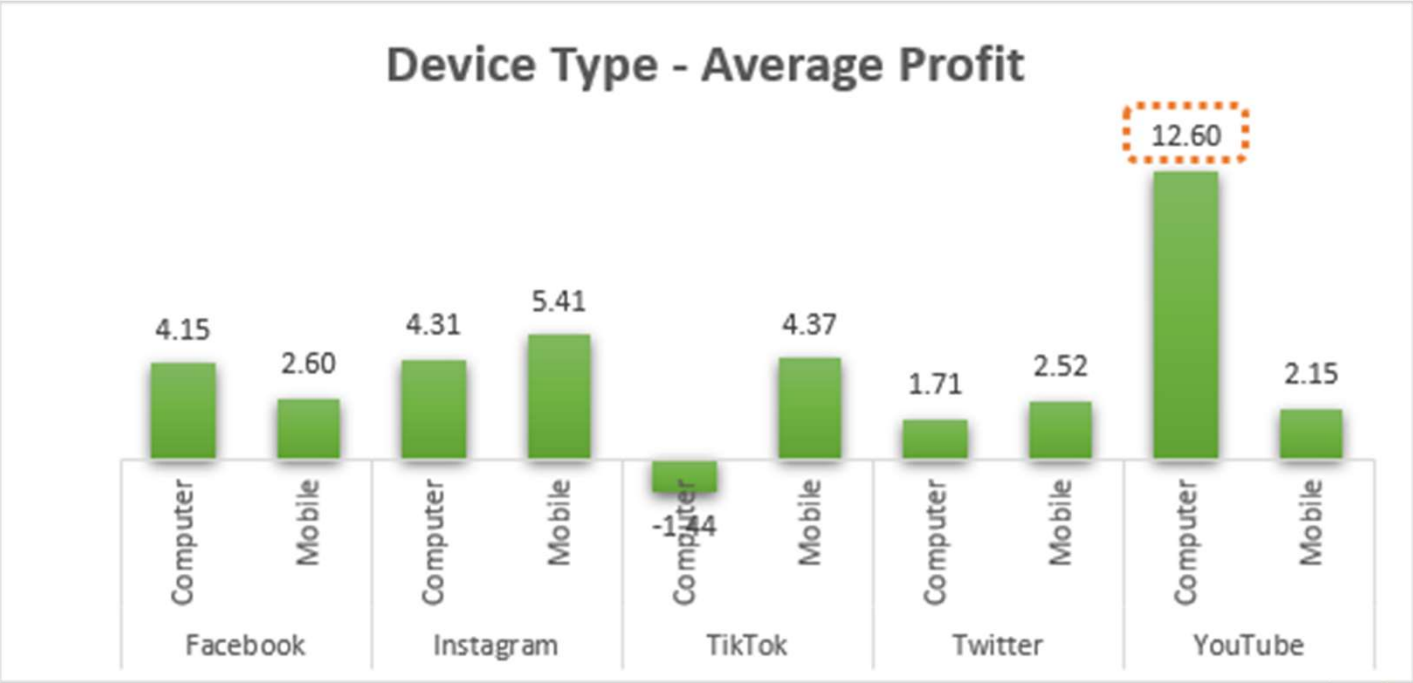
- Clearly, YouTube records the highest profits across both new and old customers.

Impact of season on profit

```
        Pairwise comparisons using t tests with pooled SD

data:  data$Profit and data$season

       fall spring summer
spring 1.00 -      -
summer 1.00 1.00   -
winter 1.00 1.00   0.87

P value adjustment method: bonferroni

  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$g
                    diff    lwr.ci   upr.ci   pval
spring-fall   -1.0331162 -5.483518 3.417285 0.9351
summer-fall   -1.8550271 -5.952206 2.242152 0.6564
winter-fall    0.1943167 -3.992805 4.381439 0.9994
summer-spring -0.8219109 -5.040606 3.396784 0.9602
winter-spring  1.2274329 -3.078668 5.533534 0.8872
winter-summer  2.0493438 -1.890619 5.989306 0.5463

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

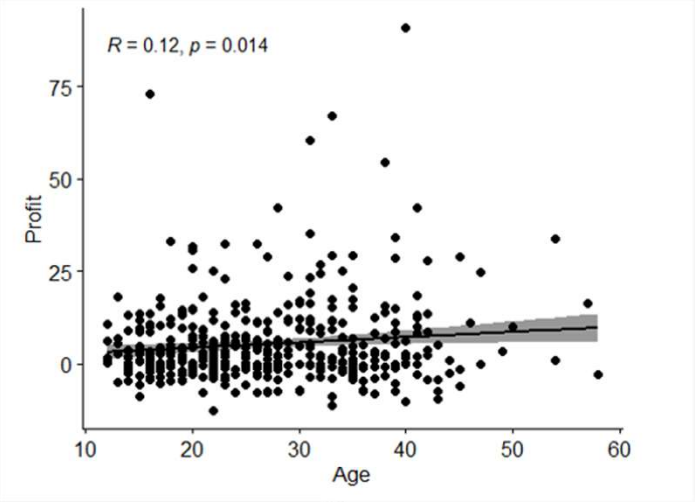
- According to the pairwise T-test using Bonferroni and Scheffe methods, All of the season pairs have equal average profit.
- Thus, the choice of the season period doesn't significantly impact profit.

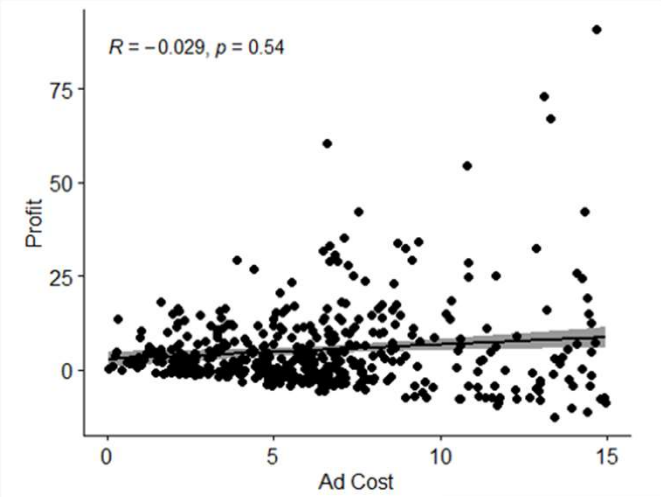
Impact of device type on profit



- Overall average profit for Computer devices Is greater than for Mobilephones.
- This is mostly driven by YouTube users (16%) generating maximum profit i.e. 12.6.
- For our use case, given an investment, there is no control in the selection of device type for the target user.

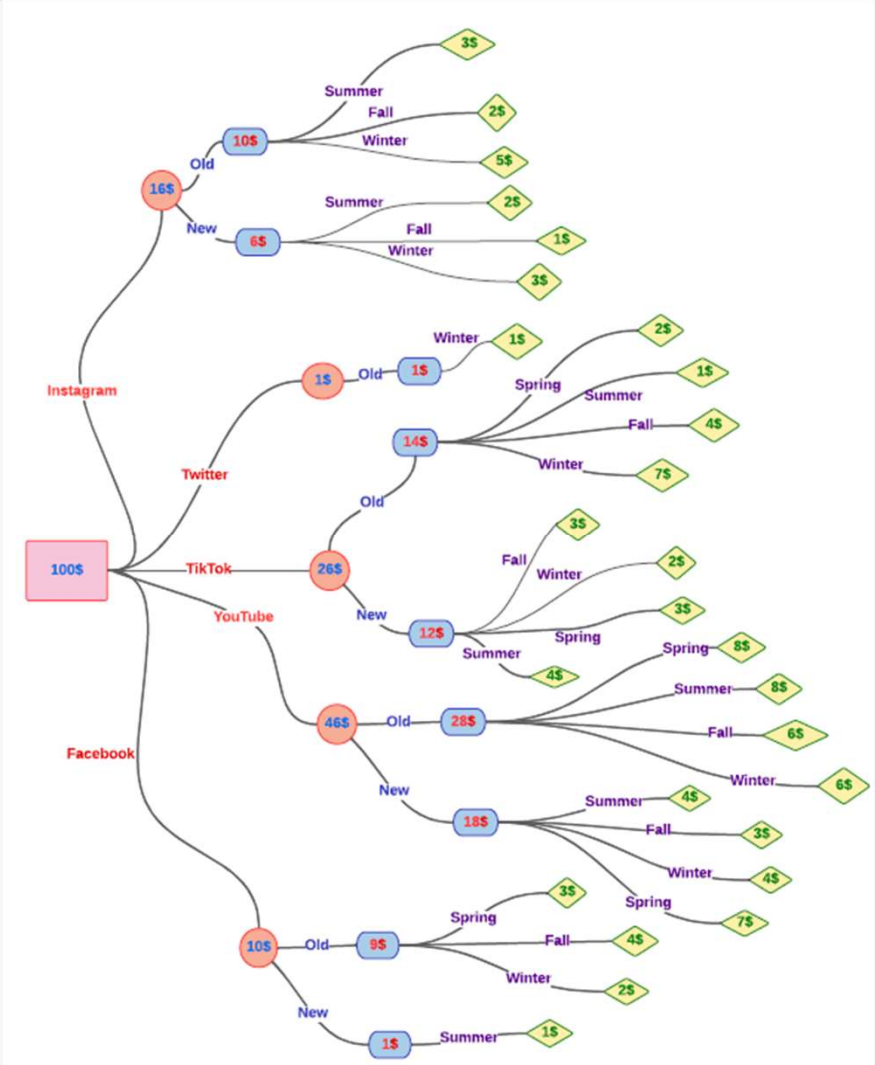
Impact of age and Ad cost on profit






```
##
## Call:
## lm(formula = profit ~ adcost + age, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.935  -6.077  -2.018   3.207  81.408
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.45097    1.68225   0.268   0.7888
## adcost       0.31223    0.15390   2.029   0.0431 *
## age          0.10610    0.06043   1.756   0.0798 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.98 on 443 degrees of freedom
## Multiple R-squared:  0.02259,    Adjusted R-squared:  0.01817
## F-statistic: 5.119 on 2 and 443 DF,  p-value: 0.006343
```

- The correlation between Age Vs. Profit and Ad Cost Vs. Profit are insignificant.
- According to the linear regression analysis shown beside, neither Ad cost nor Ad cost + age impact profit.



Cost Distribution

- To conclude, we recommend investing more in YouTube and TikTok, especially focusing on existing customer engagement.