A/B Testing Analysis of Digital Marketing Campaigns

Berkay CAYAN

```
library(readr)
data <- read_delim("marketing_AB.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE)
```


Summary

In this project, an analysis based on A/B testing will be conducted to evaluate the effectiveness of digital marketing campaigns. We have a dataset where users are assigned to different advertisement groups, and their conversion status is tracked. The data includes information such as the advertisement group the users belong to (test_group), the total number of ad impressions (total_ads), the day with the highest ad impressions (most_ads_day), and the hour with the highest ad impressions (most_ads_hour). The objective is to understand the effects of ad impressions on conversion rates and identify the days and time slots with the highest conversion rates.

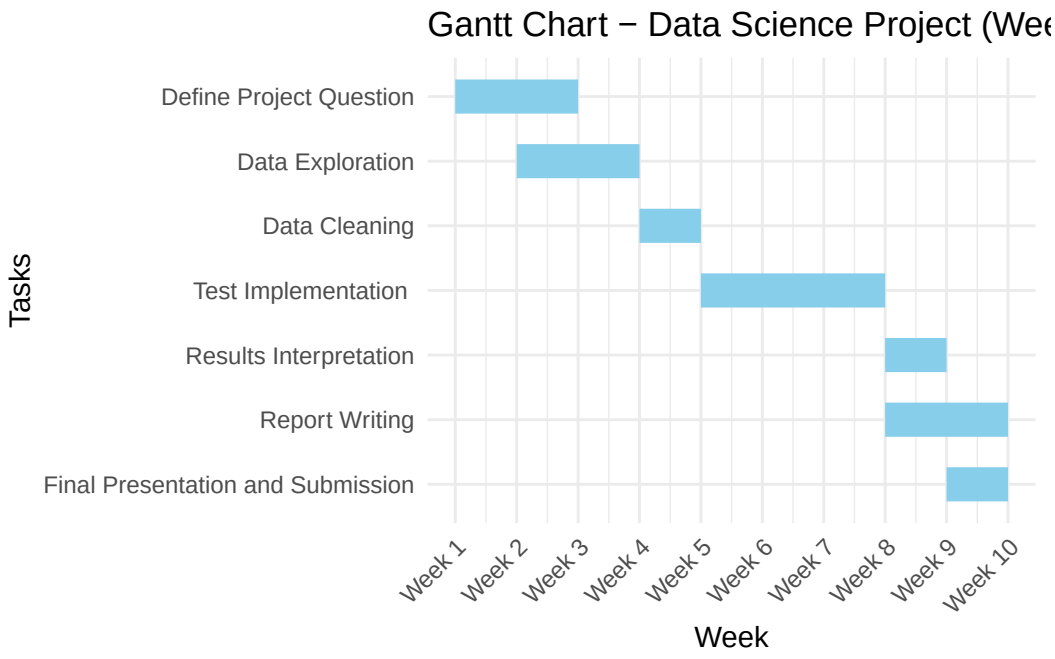
Problem:

To understand how ad impressions impact user conversion rates. To answer this question, various analyses will be conducted. Users are divided into two different test groups (ad and psa), and conversion rates will be compared for each group. Additionally, the relationship between the timing of ad impressions (the day and hour with the highest impressions) and conversion rates will be examined.

```
tasks <- data.frame(
  task = c("Define Project Question", "Data Exploration", "Data Cleaning",
           "Test Implementation ",
           "Results Interpretation", "Report Writing", "Final Presentation and Submission"),
  start = c(1, 2, 4, 5, 8, 8, 9),
```

```
  end = c(3, 4, 5, 8, 9, 10, 10)
)


tasks$task <- factor(tasks$task, levels = rev(tasks$task[order(tasks$start)]))
ggplot(tasks, aes(x = start, xend = end, y = task, yend = task)) +
  geom_segment(size = 6, color = "skyblue") +
  scale_x_continuous(name = "Week", breaks = 1:10, labels = paste("Week", 1:10)) +
  labs(title = "Gantt Chart - Data Science Project (Week 1 to Week 10)", y = "Tasks") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



Exploratory Data Analysis:

```
colSums(is.na(data))
```

```
         ...1       user_id    test_group     converted     total_ads
            0             0             0             0             0
 most_ads_day most_ads_hour
            0             0
```

```
# If there are missing values, check which columns have them
anyNA(data)
```

```
[1] FALSE
```

```
# Check the data types of the columns in the dataset
sapply(data, class)
```

```
         ...1       user_id    test_group     converted     total_ads
    "numeric"     "numeric"   "character"     "logical"     "numeric"
 most_ads_day most_ads_hour
  "character"     "numeric"
```

```
ad_group <- data %>% filter(test_group == "ad") %>% pull(converted)
psa_group <- data %>% filter(test_group == "psa") %>% pull(converted)

ad_group <- as.numeric(data$converted[data$test_group == "ad"])
psa_group <- as.numeric(data$converted[data$test_group == "psa"])
```

The “ad” group consists of users who have seen the advertisements. The “psa” group consists of users who have not seen the advertisements.

The purpose of this test is to evaluate the effectiveness of advertisements by measuring the difference in conversion rates between the users who saw the ads (ad) and those who did not (psa).

```
cat("total_ads\n")
```

```
total_ads
```

```
summary(data$total_ads)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    4.00   13.00   24.82   27.00 2065.00
```

```r
cat("\n")
```

```r
cat("most_ads_hour\n")
```

```
most_ads_hour
```

```r
summary(data$most_ads_hour)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   11.00   14.00   14.47   18.00   23.00
```

```r
ctr_day <- prop.table(table(data$most_ads_day, data$converted), margin = 1)


ctr_day_sorted <- ctr_day[order(ctr_day[, 2], decreasing = TRUE), ]
print(ctr_day_sorted)
```

```
                FALSE       TRUE
  Monday    0.96718845 0.03281155
  Tuesday   0.97015966 0.02984034
  Wednesday 0.97505809 0.02494191
  Sunday    0.97552435 0.02447565
  Friday    0.97778810 0.02221190
  Thursday  0.97842906 0.02157094
  Saturday  0.97894930 0.02105070
```

```r
highest_CVR_day <- rownames(ctr_day_sorted)[1]
cat("The highest CVR is on", highest_CVR_day, "\n")
```

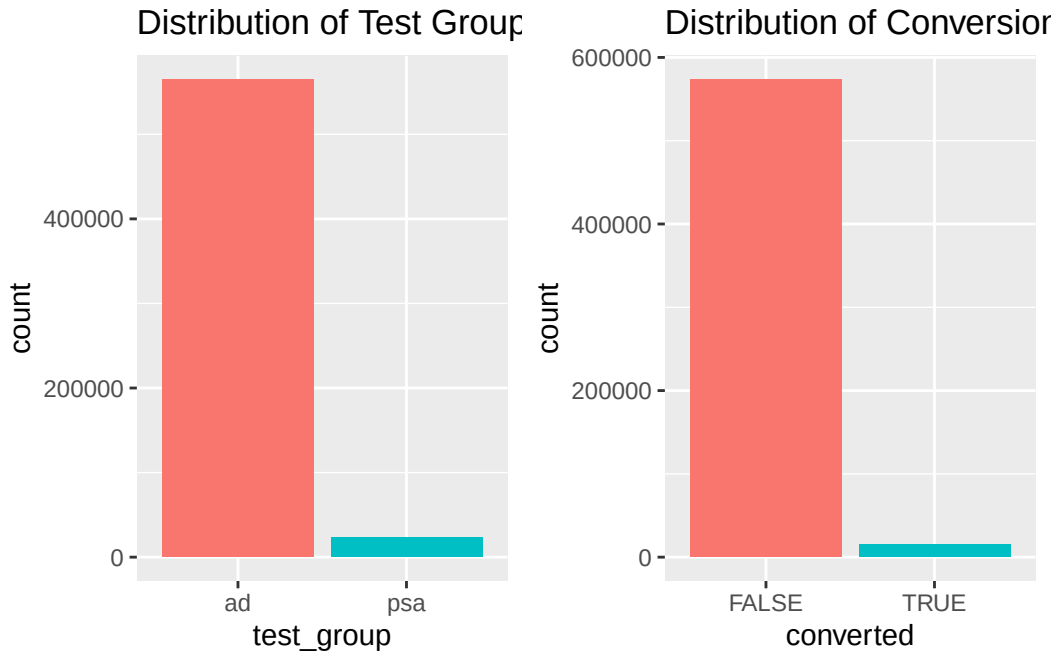
```
The highest CVR is on Monday
```

In this analysis, the relationship between the most_ads_day variable and the converted variable was examined, and conversion rates (CVR) were calculated for each day. According to the results:

- The highest conversion rate occurred on Monday (3.28%).
- Compared to other days, Monday stands out as a more effective day for conversions.
- This finding suggests that focusing on Monday in marketing strategies could be beneficial for increasing conversion rates.

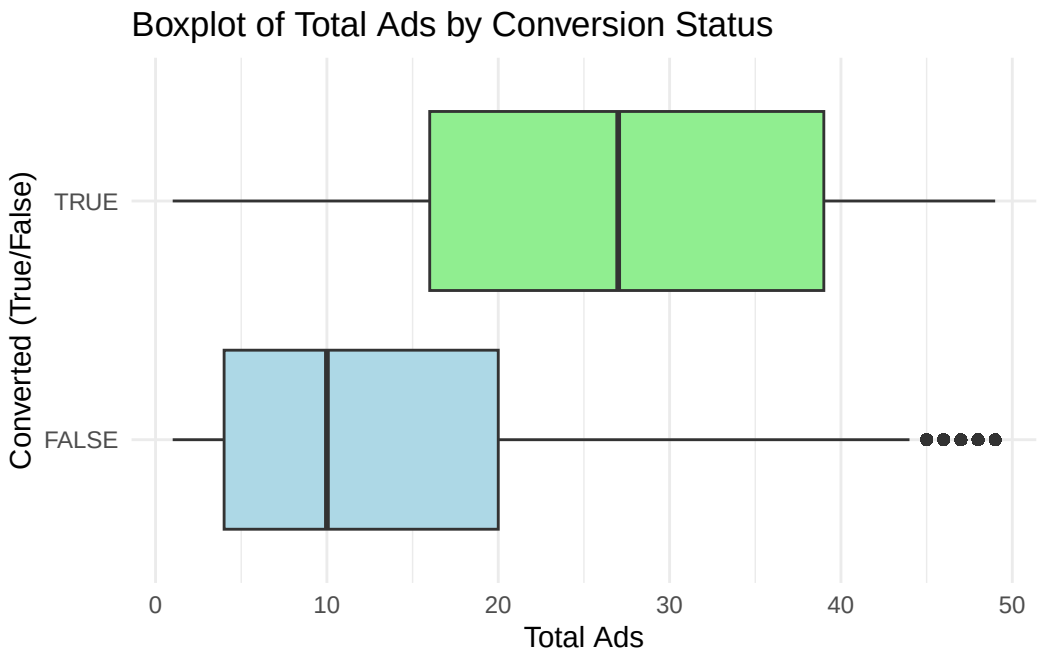
Visualization:

```
options(scipen = 999)

plot1 <- ggplot(data, aes(x = test_group, fill = test_group)) +
  geom_bar() +
  labs(title = "Distribution of Test Groups") +
   theme(legend.position = "none")

plot2 <- ggplot(data, aes(x = converted, fill = as.factor(converted))) +
  geom_bar() +
  labs(title = "Distribution of Conversion Status") +
   theme(legend.position = "none")


grid.arrange(plot1, plot2, ncol = 2)
```

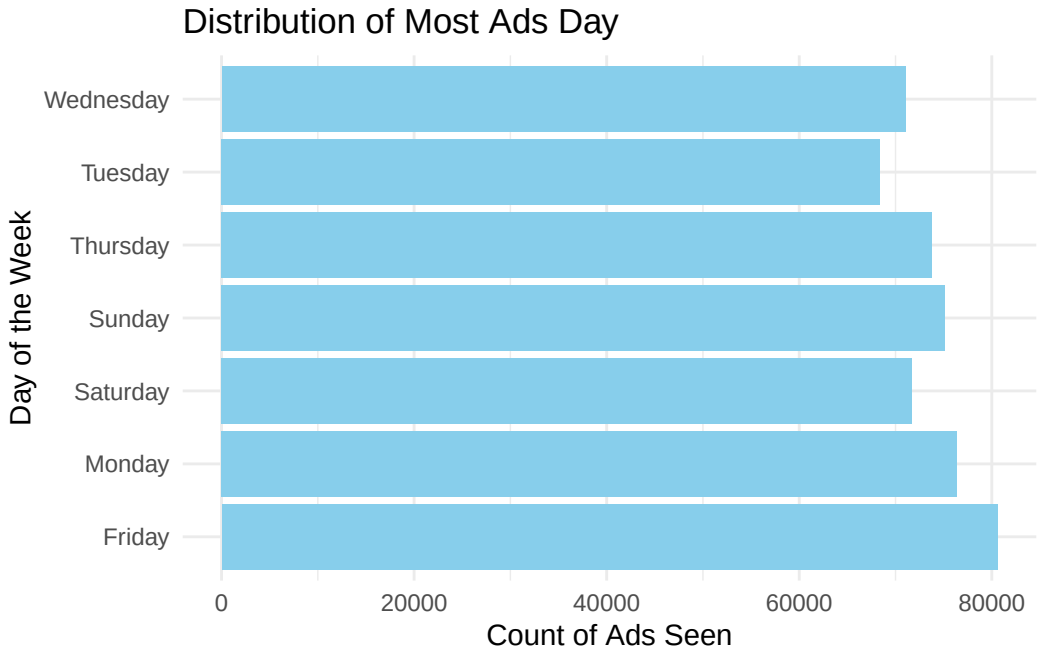



```
filtered_data <- data %>% filter(total_ads < 50)


ggplot(filtered_data, aes(x = total_ads, y = as.factor(converted), fill = as.factor(converte
  geom_boxplot() +
```

```
  labs(title = "Boxplot of Total Ads by Conversion Status",
       x = "Total Ads",
       y = "Converted (True/False)") +
  theme_minimal() +
  scale_fill_manual(values = c("lightblue", "lightgreen"))+
  theme(legend.position = "none")
```




```
ggplot(filtered_data, aes(x = most_ads_day)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Distribution of Most Ads Day",
       x = "Day of the Week",
       y = "Count of Ads Seen") +
  theme_minimal() +
  coord_flip()
```



The conversion rate is a critical metric that measures the success of an advertising campaign and evaluates the effectiveness of marketing strategies. Since it shows the ratio of ad impressions or interactions that lead to conversions, it plays an important role in the decision-making processes of advertisers, digital marketing teams, and strategists.

$$\text{Conversion Rate} = \left(\frac{\text{Total Conversions}}{\text{Total Ads or Impressions}} \right) \times 100$$

Figure 1: CR Calculation

```
conversion_rate_data <- data %>%
  group_by(test_group) %>%
  summarise(
    total_clicks = sum(converted, na.rm = TRUE),
    total_ads = sum(total_ads, na.rm = TRUE),
    conversion_rate = (total_clicks / total_ads) * 100
  )


ggplot(conversion_rate_data, aes(x = test_group, y = conversion_rate, fill = test_group)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
```

```
  labs(title = "Conversion Rate by Test Group",
       x = "Test Group",
       y = "Conversion Rate (%)") +
  theme_minimal() +
  geom_text(aes(label = round(conversion_rate, 2)), vjust = -0.5, size = 5)
```

- Ad Group (CR = 0.1): This means that 10% of the users in the ad group completed the desired action (e.g., clicking, making a purchase, signing up) after being exposed to the advertisement.
- PSA Group (CR = 0.07): This means that 7% of the users in the PSA group took the desired action after seeing the content related to the Public Service Announcement (PSA), which typically doesn't include paid ads.

Statistical Testing:

Normality assumption

Many parametric tests are based on the assumption of normal distribution. For these tests to be valid, the data must follow a normal distribution. If the data does not follow a normal distribution, the results may be misleading.

To perform parametric tests, it is important to first check if the data meets the assumption of normality. If the data does not follow a normal distribution, the results of parametric tests may be misleading. To assess whether the data follows a normal distribution, various normality tests can be used. These include: - Shapiro-Wilk Test - Kolmogorov-Smirnov Test: - Anderson-Darling Test:

Since our dataset contains more than 5,000 observations, it is not possible to apply the Shapiro-Wilk test. Therefore, we will use the Kolmogorov-Smirnov test.

Kolmogorov-Smirnov test

ad_group: Null Hypothesis (H): : The distribution of the data is normal. (The data follows a normal distribution.)

Alternative Hypothesis (H): The distribution of the data is not normal. (The data does not follow a normal distribution.)

psa_group:

Null Hypothesis (H): The distribution of the data is normal. Alternative Hypothesis (H): The distribution of the data is not normal.

```
options(scipen = 999)
data$converted <- as.numeric(data$converted)
data$test_group <- as.factor(data$test_group)


ad_group <- na.omit(ad_group)
psa_group <- na.omit(psa_group)

# Kolmogorov-Smirnov
ks_test_ad_group <- ks.test(ad_group, "pnorm", mean(ad_group), sd(ad_group))
ks_test_psa_group <- ks.test(psa_group, "pnorm", mean(psa_group), sd(psa_group))


print(ks_test_ad_group)
```

```
    Asymptotic one-sample Kolmogorov-Smirnov test

data:  ad_group
D = 0.53877, p-value < 0.00000000000000022
alternative hypothesis: two-sided
```

```
print(ks_test_psa_group)
```

```
    Asymptotic one-sample Kolmogorov-Smirnov test

data:  psa_group
D = 0.53577, p-value < 0.00000000000000022
alternative hypothesis: two-sided
```

Both ad_group and psa_group do not follow a normal distribution according to the Kolmogorov-Smirnov test. Since the data fails the normality test, parametric tests that assume normality (such as t-tests or ANOVA) would not be appropriate for analyzing these data. Instead, non-parametric tests, which do not assume normality, should be considered for further analysis.

Mann-Whitney U test

Null Hypothesis (H): There is no significant difference in conversion rates (converted) between the "ad" and "psa" groups.

Alternative Hypothesis (H): There is a significant difference in conversion rates between the "ad" and "psa" groups.

```
mann_whitney_test <- wilcox.test(converted ~ test_group, data = data)


print(mann_whitney_test)
```

```
    Wilcoxon rank sum test with continuity correction

data:  converted by test_group
W = 6691636830, p-value = 0.0000000000001705
alternative hypothesis: true location shift is not equal to 0
```

The Mann-Whitney U test tests the median differences between groups. Since the p-value is very small, the null hypothesis is rejected, and it is concluded that there is a significant difference in conversion rates between the "ad" and "psa" groups.

Results

According to the results of the A/B test, the conversion rate of the advertising group (10%) is higher, meaning that users exposed to ads have a significantly higher conversion rate compared to the group that did not see ads. It was concluded that ad impressions have a significant impact on conversion rates, and marketing strategies can be shaped accordingly. It was suggested that focusing on Monday, the day with the highest conversion rate, could be more effective in marketing strategies. In this process, non-parametric tests were used instead of parametric tests to obtain accurate and reliable results.