Aryan Sharma

Advance Data Science
Difference in Difference Analysis

1. Introduction

This analysis investigates the impact of a policy intervention introduced in November 1930 on the number of banks that remained open in two districts, District 6 and District 8. The analysis uses the Difference-in-Differences (DiD) methodology to assess the policy's effects by comparing the number of banks before and after the intervention across the two districts.

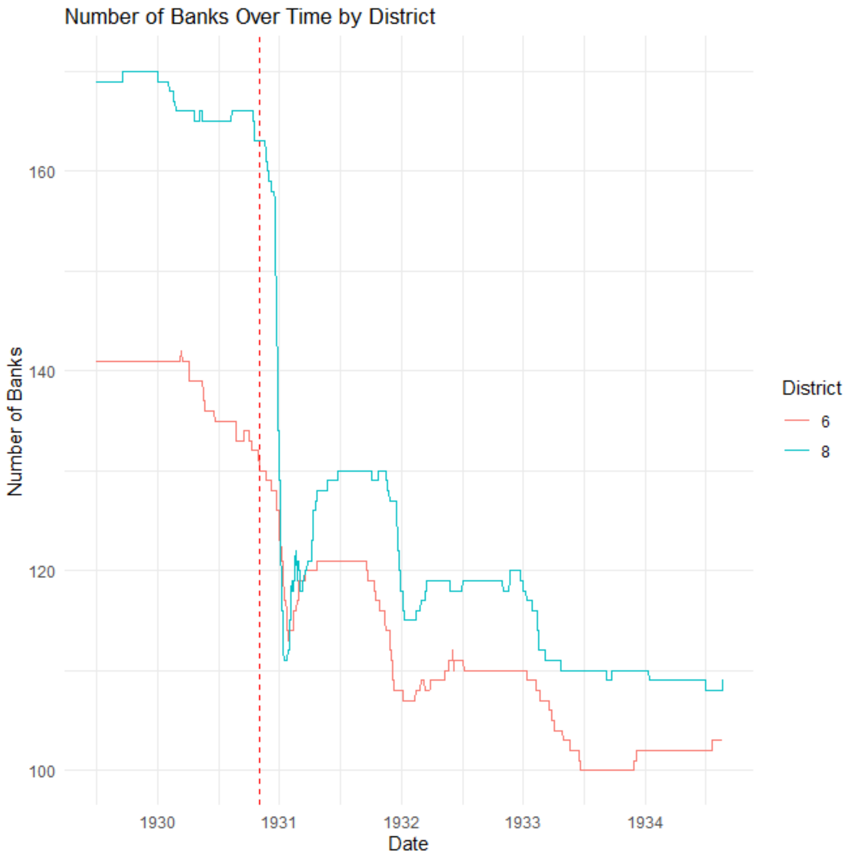
2. Dataset Description

The dataset consists of daily records of the number of banks open in two districts (District 6 and District 8) from July 1929 to December 1934. Descriptive Analysis is as follows:

district	after_policy	mean_banks	median_banks	sd_banks	min_banks	max_banks
6	0	140	141	2.93	131	141
6	1	114	110	12.3	100	142
8	0	169	169	1.77	163	170
8	1	125	119	19.6	108	169

3. Objective of the Study

The primary objective of this analysis is to assess the effect of the policy intervention on the number of banks open in the two districts using a Difference-in-Differences (DiD) approach. The key question is whether the policy had a differential impact across the two districts.



4. Methodology and DiD Model Explanation

The Difference-in-Differences (DiD) methodology is used to estimate the impact of the policy by comparing the change in outcomes (number of banks open) in the treatment group (District 8) to the change in outcomes in the control group (District 6) before and after the policy intervention.
The DiD model used is specified as follows:

*did_model <- lm(banks ~ after_policy * district, data = data_long)*

Where:
- banks: Number of banks open
- after_policy: 1 if after November 1930, 0 if before
- district: 6 for District 6 and 8 for District 8
- after_policy * district: Interaction term capturing the DiD effect

5. Summary of Results and Descriptive Analysis

```
lm(formula = banks ~ after_policy * district, data = data_long)

Residuals:

   Min      1Q  Median      3Q     Max

-17.456 -11.697  -3.697   4.544  43.544

Coefficients:

                    Estimate Std. Error t value Pr(>|t|)

(Intercept)          52.9919     4.8811  10.857  < 2e-16 ***

after_policy         25.4272     5.2360   4.856 1.25e-06 ***

district             14.4837     0.6903  20.982  < 2e-16 ***

after_policy:district  -8.6041    0.7405 -11.620  < 2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.31 on 3752 degrees of freedom

Multiple R-squared:  0.4548,      Adjusted R-squared:  0.4544

F-statistic:  1043 on 3 and 3752 DF,  p-value: < 2.2e-16
```

The results from the DiD model reveal the following insights:
- The baseline number of banks in District 6 before the policy was approximately 53 banks.
- After the policy, the number of banks in District 6 increased by about 25.43 banks.
- District 8 had around 14.48 more banks than District 6 before the intervention.
- The interaction term, which captures the DiD effect, suggests that the policy caused a decrease of about 8.60 banks in District 8 relative to District 6.

6. Interpretation and Implications

The policy intervention had a positive effect on the number of banks in District 6, but it appears to have had a negative impact on District 8, where the number of banks declined by approximately 8.60 banks relative to District 6. This suggests that the policy intervention may have inadvertently caused adverse effects in District 8, potentially due to differences in economic conditions, regulatory enforcement, or other unobserved factors.

7. Conclusion

The Difference-in-Differences analysis provides robust evidence that the policy intervention had heterogeneous effects across the two districts. While District 6 experienced a net positive impact, District 8 experienced a relative decline in the number of banks open, indicating that further investigation is required to understand the underlying reasons for these divergent effects.

Part - 2

1. Propensity Score Matching Analysis of Groupon Deals

The primary goal of this analysis is to determine whether having a minimum requirement that is as a minimum number of committed buyers for Groupon deals impacts the outcomes of these deals. Specifically, the outcomes of interest are revenue, quantity sold, and Facebook likes received. To assess this, deals are divided into two groups: the control group (deals without the minimum requirement) and the treatment group (deals with the minimum requirement). Propensity score matching is used to analyze the effect of this minimum requirement.

2. Load the data into your environment and perform any necessary cleaning steps

The dataset was loaded from an Excel file and read into the R environment. Basic cleaning steps included checking for missing values and ensuring that data types were correctly formatted.

```
file_path <- ("C:/Users/91884/Desktop/BAIS/Advance data science/Assignment 2/groupon.xlsx")
data <- read_excel(file_path, sheet = "groupon")
summary(data)
str(data)
```

3. Explain what a propensity score is and how it is used in the context of this Groupon dataset

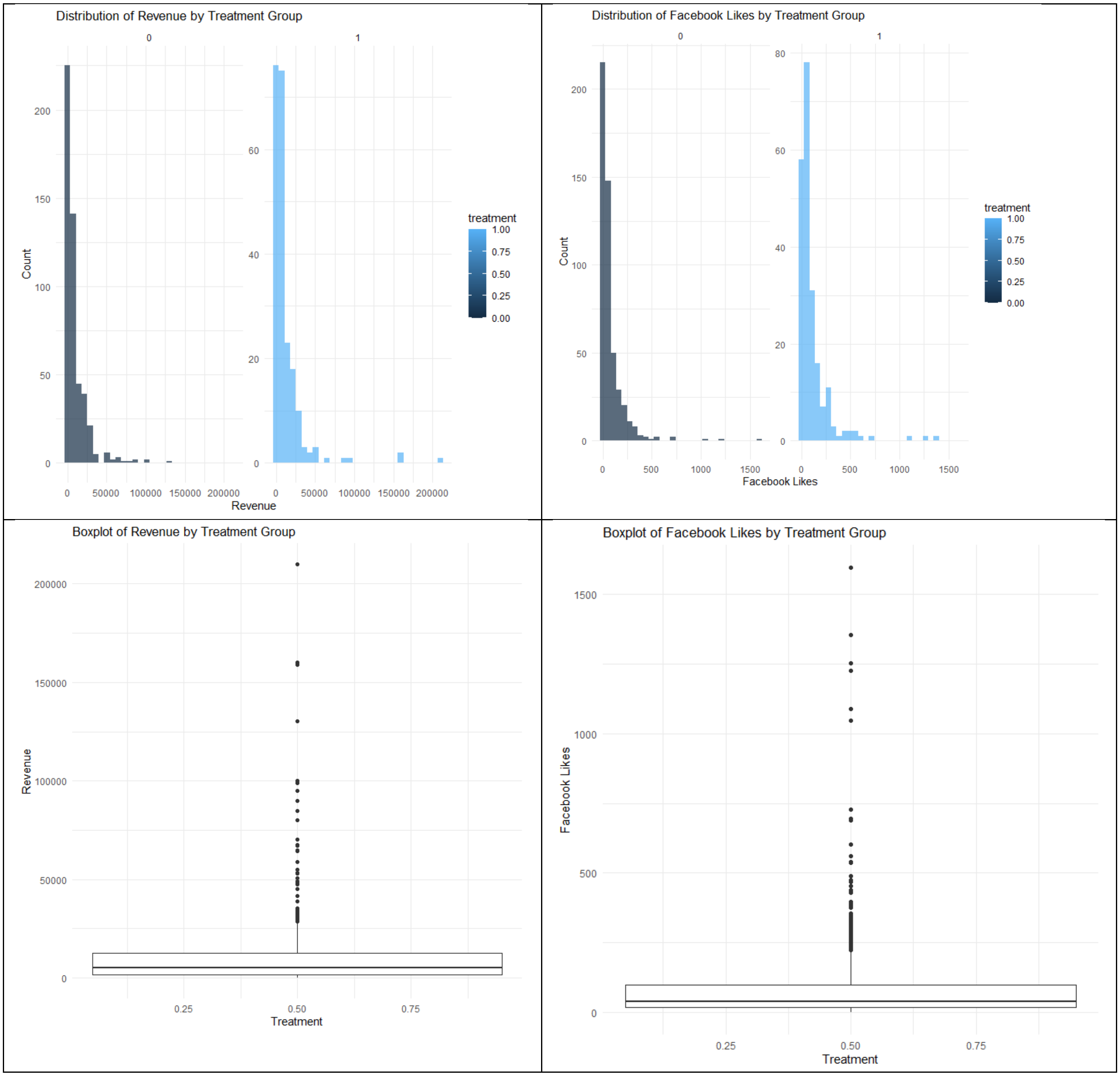
A propensity score is the probability of an assignment to a treatment group given a set of observed covariates. In this dataset, the propensity score helps in matching deals with and without minimum requirements to minimize bias and ensure that the groups being compared are similar. Calculating propensity scores before performing matching is essential to reduce selection bias and account for potential confounding variables.

```
data$propensity_score <- predict(propensity_model, type = "response")
head(data)
```

Aryan Sharma

4. Visualize the distribution between treatment and control groups

Visualizations including histograms, boxplots, and density plots were generated to explore the distribution of revenue and Facebook likes across treatment and control groups. These visualizations help identify differences in distributions and potential outliers.

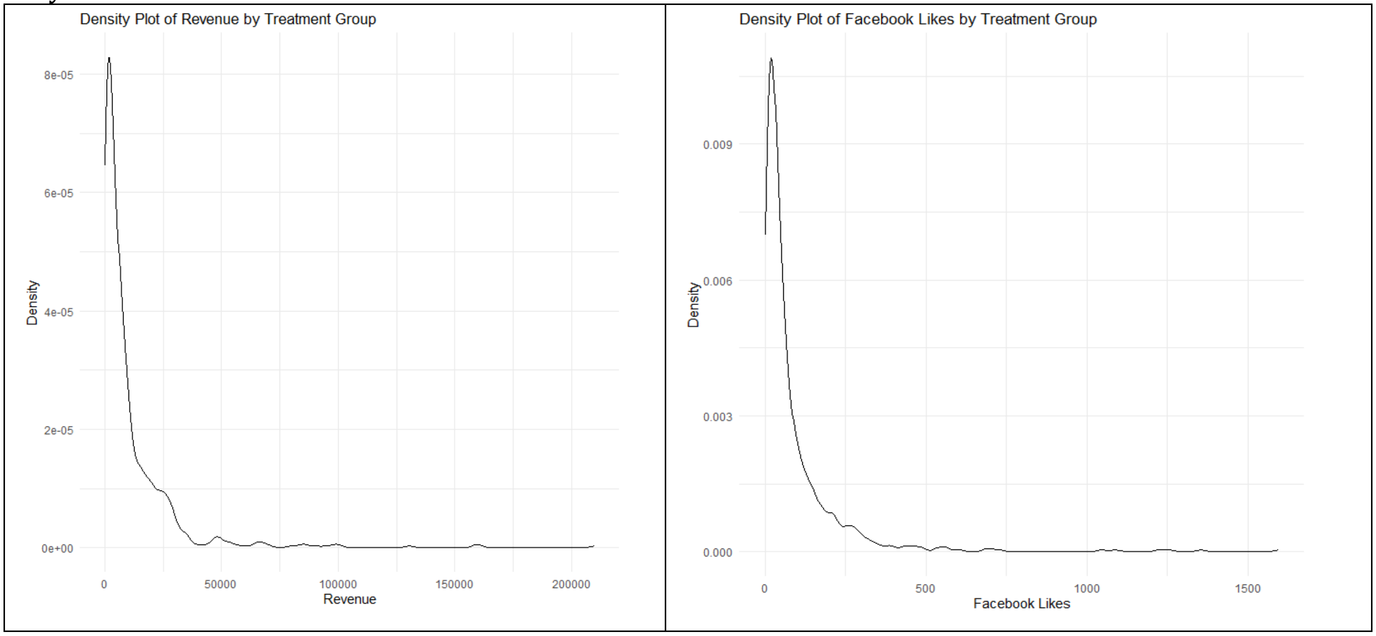


Distribution of Revenue by Treatment Group

Distribution of Facebook Likes by Treatment Group

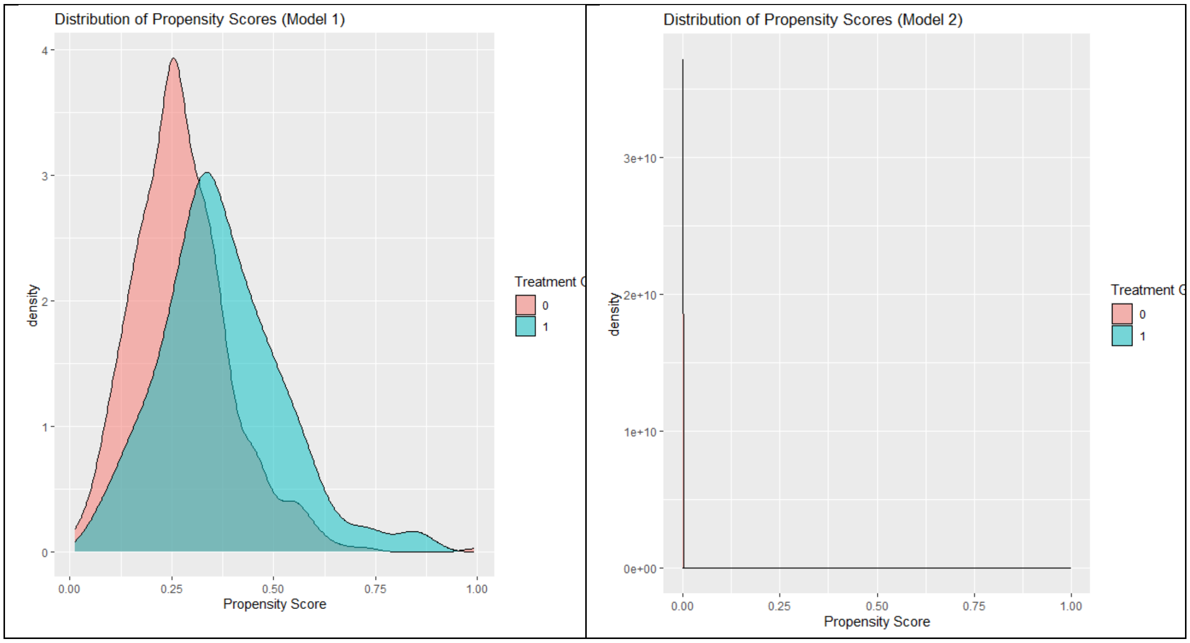
Boxplot of Revenue by Treatment Group

Boxplot of Facebook Likes by Treatment Group

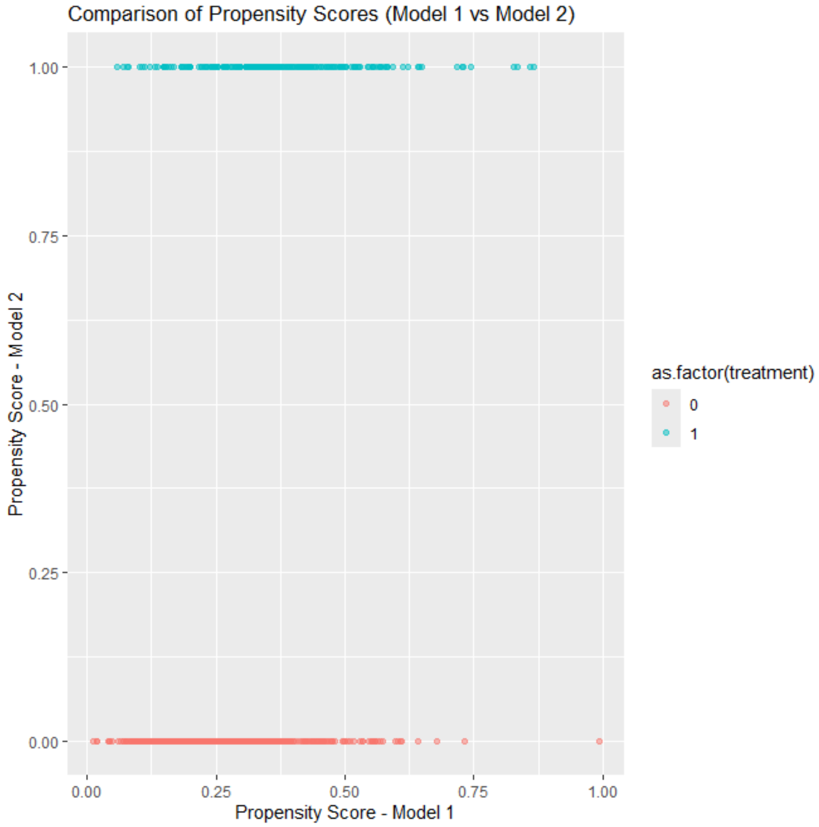


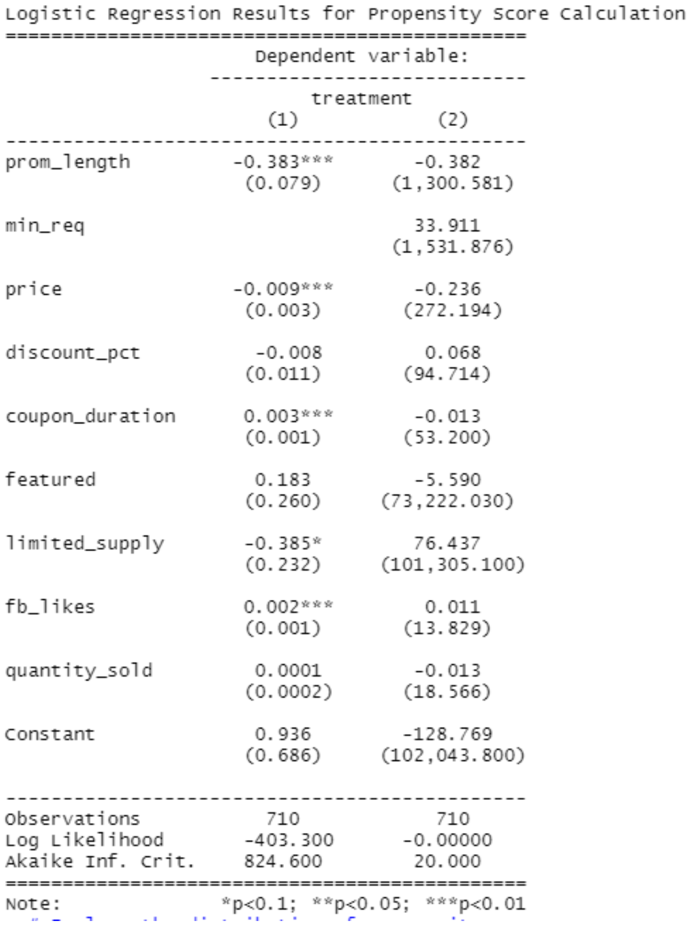
5. Calculate a propensity score using a regression model

To calculate the propensity score, two logistic regression models were used. The models included relevant covariates such as promotion length, price, discount percentage, coupon duration, and whether the deal was featured or had limited supply. The inclusion of 'min_req' in the second model resulted in differences in propensity scores.



Aryan Sharma

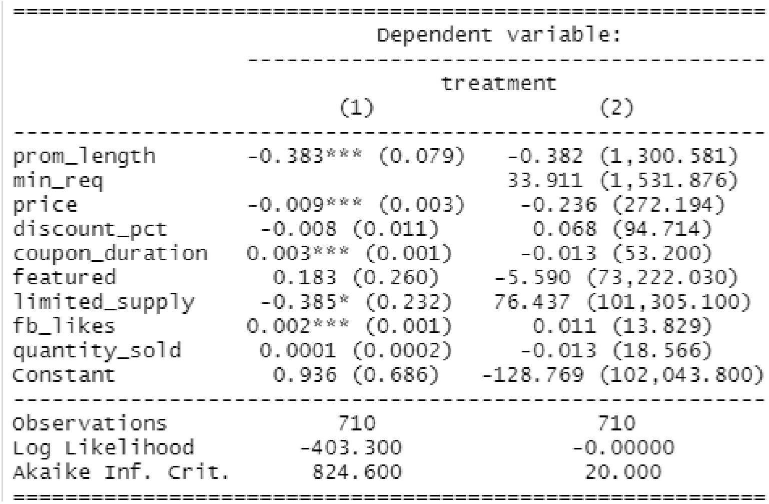




```
Logistic Regression Results for Propensity Score Calculation
=============================================
                      Dependent variable:
                  ---------------------------
                           treatment
                      (1)            (2)
---------------------------------------------
prom_length        -0.383***       -0.382
                   (0.079)      (1,300.581)

min_req                            33.911
                                (1,531.876)

price              -0.009***       -0.236
                   (0.003)       (272.194)

discount_pct        -0.008          0.068
                   (0.011)        (94.714)

coupon_duration    0.003***        -0.013
                   (0.001)        (53.200)

featured            0.183          -5.590
                   (0.260)      (73,222.030)

limited_supply     -0.385*         76.437
                   (0.232)     (101,305.100)

fb_likes           0.002***         0.011
                   (0.001)        (13.829)

quantity_sold       0.0001         -0.013
                   (0.0002)       (18.566)

Constant            0.936         -128.769
                   (0.686)     (102,043.800)

---------------------------------------------
Observations         710            710
Log Likelihood     -403.300       -0.00000
Akaike Inf. Crit.  824.600         20.000
=============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

6. Which variables in the above regression model were significant?

The variables that were significant in predicting whether a deal belonged to the treatment group included promotion length, price, and Facebook likes. These variables were important as they influenced the likelihood of a deal having a minimum requirement, impacting the propensity score.



```
==========================================================
                            Dependent variable:
                  ----------------------------------------
                                treatment
                        (1)                   (2)
----------------------------------------------------------
prom_length       -0.383*** (0.079)    -0.382 (1,300.581)
min_req                                33.911 (1,531.876)
price             -0.009*** (0.003)     -0.236 (272.194)
discount_pct       -0.008 (0.011)        0.068 (94.714)
coupon_duration   0.003*** (0.001)      -0.013 (53.200)
featured            0.183 (0.260)     -5.590 (73,222.030)
limited_supply     -0.385* (0.232)    76.437 (101,305.100)
fb_likes          0.002*** (0.001)       0.011 (13.829)
quantity_sold      0.0001 (0.0002)      -0.013 (18.566)
Constant            0.936 (0.686)   -128.769 (102,043.800)
----------------------------------------------------------
Observations             710                   710
Log Likelihood        -403.300              -0.00000
Akaike Inf. Crit.      824.600               20.000
==========================================================
```

In **Model 1**, three variables were statistically significant in predicting whether a Groupon deal belonged to the treatment group:

1. **prom_length**: The coefficient was -0.383 with a standard error of 0.079, and it was statistically significant at the 0.001 level. This suggests that longer promotion lengths were associated with a lower likelihood of the deal belonging to the treatment group.

2. **price**: The coefficient was -0.009 with a standard error of 0.003, and it was statistically significant at the 0.001 level. Lower prices were associated with a higher likelihood of being in the treatment group.

3. **fb_likes**: The coefficient was 0.002 with a standard error of 0.001, and it was statistically significant at the 0.001 level. Deals with higher Facebook likes were more likely to be part of the treatment group.

4. **limited_supply**: This variable was marginally significant with a p-value slightly below 0.05. The coefficient was -0.385 with a standard error of 0.232, indicating that deals with limited supply were slightly less likely to be in the treatment group.

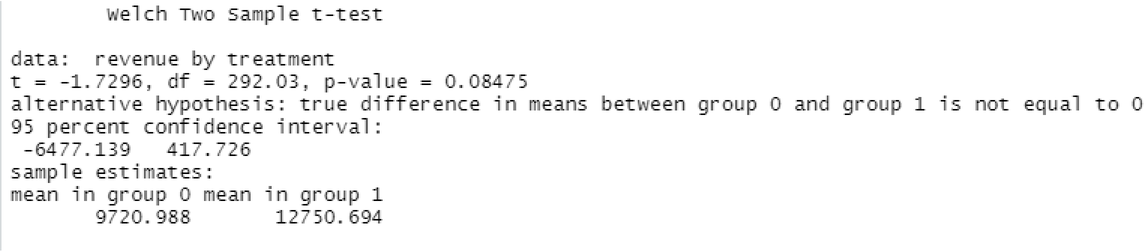
In **Model 2**, after including min_req as an additional covariate, none of the variables were statistically significant. The inclusion of min_req resulted in extremely large standard errors and unstable parameter estimates, suggesting the presence of multicollinearity or model misspecification. For instance:

- min_req had a coefficient of 33.911 with a standard error of 1,531.876, making the estimate highly unreliable.
- Other covariates such as featured, limited_supply, and quantity_sold had similarly large standard errors, indicating that the results from Model 2 were not interpretable.

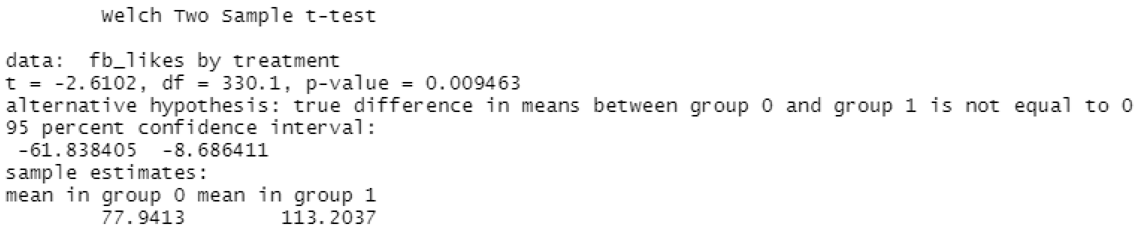
7. Conduct two-sample t-tests for 'revenue' and 'fb_likes'

Two-sample t-tests were conducted for 'revenue' and 'fb_likes' to compare the means between the treatment and control groups. The p-values and confidence intervals were examined to determine whether the differences between the groups were statistically significant.

```
t_test_revenue <- t.test(revenue ~ treatment, data = data)
t_test_fb_likes <- t.test(fb_likes ~ treatment, data = data)
```



```
        Welch Two Sample t-test

data:  revenue by treatment
t = -1.7296, df = 292.03, p-value = 0.08475
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -6477.139   417.726
sample estimates:
mean in group 0 mean in group 1
       9720.988       12750.694
```



```
        Welch Two Sample t-test

data:  fb_likes by treatment
t = -2.6102, df = 330.1, p-value = 0.009463
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -61.838405  -8.686411
sample estimates:
mean in group 0 mean in group 1
        77.9413        113.2037
```

8. What issues arose when 'min_req' was included in the logistic regression model?

Including 'min_req' in the logistic regression model for calculating propensity scores created multicollinearity and made it difficult to match treated and control units. This issue complicated the matching process and introduced bias in the estimation of treatment effects.