Name: Isaac Ng Sheng
Matriculation Number: A0258037B
Course: BT4241 Causal Impact Analysis
Assignment 1

Q1 Concepts and Theories

1.1 Potential Outcome Model

Consider a population of 6 units:

i	T_i	$Y_i(1)$	$Y_i(0)$
1	0	1	1
2	0	0	0
3	0	0	1
4	1	1	0
5	1	1	0
6	1	1	1

Q1) What is the average treatment effect (ATE)?

$$\begin{aligned} ATE &= E\left[Y_i(1) - Y_i(0)\right] \\ &= E\left[Y_i(1)\right] - E\left[Y_i(0)\right]. \end{aligned}$$

ATE = [0 + 0 + (-1) + 1 + 1 + 0] / 6 = ⅙ = 0.167 (3s.f.)

Q2) What is the average treatment effect of the treated (ATT)?

$$\begin{aligned} \text{ATT} &= E\left[Y_i(1) - Y_i(0) \mid T_i = 1\right] \\ &= E\left[Y_i(1) \mid T_i = 1\right] - E\left[Y_i(0) \mid T_i = 1\right]. \end{aligned}$$

ATT = [1 + 1 + 0] / 3 = ⅔ = 0.667 (3s.f.)

Q3)

a) No, this does not measure ATE. This equation measures the difference between the two averages of potential outcomes under the condition that only considers those that are treated.
   ATE includes both those that are treated and untreated (the entire population) while this equation does not.
   Overall, it measures the average treated effect of the treated (ATT) instead. This ATT can be observed as there is a restriction, ignoring the counterfactual outcome of Y(0), that cannot be observed if outcome Y(1) is observed.
b) Yes, this does measure ATE, it averages out the potential outcomes before finding the difference. With ATE, it compares the entire population (entire dataset) which this equation does.
   However, this cannot be observed as at any point in time, we cannot observe both outcomes for any individual. (Whether each individual receives the treatment or not).

It can only be estimated through experimental randomization or causal inference techniques like instrumental variables.

1.2 ATE

- Average treatment effect:

$$\begin{aligned} ATE &= E\left[Y_i(1) - Y_i(0)\right] \\ &= E\left[Y_i(1)\right] - E\left[Y_i(0)\right]. \end{aligned}$$

- Average effect on the treated:

$$\begin{aligned} \text{ATT} &= E\left[Y_i(1) - Y_i(0) \mid T_i = 1\right] \\ &= E\left[Y_i(1) \mid T_i = 1\right] - E\left[Y_i(0) \mid T_i = 1\right]. \end{aligned}$$

Q1) Let ATE be the difference in smoking rates between the treatment and control groups,
Smoking Rate of Treatment Group: 5 / (150 + 5) = 0.03225806
Smoking Rate of Control Group: 100 / (2000 + 100) = 0.04761904
ATE = 0.03225806 - 0.04761904 = -0.0153609 = -0.0154 (3s.f.)
Since the ATE is approximately -0.0154, the cash incentive program approximately reduces smoking rates by about 1.54% in Palo Alto.

Q2) Let ATE be the difference in smoking rates between the treatment and control groups in both cities,
Smoking Rate of Treatment Group: 305 / (305 + 650) = 0.31937
Smoking Rate of Control Group: 2000 / (2000 + 1000) = 0.333333
ATE = 0.31937 - 0.33333 = -0.0139616 = -0.0140 (3s.f.)
Since the ATE is approximately -0.0140, the cash incentive program approximately reduces smoking rates by about 1.40% in both cities.

1.3 Measure Causal Impact of KOL

Q1)
The current method of using promotion codes as a source of causal impact of the youtube ads may be biased as there maybe problems with sample selection.
Not all users who use the promo code may have watched the advertisement as the promo code itself could be spread through word of mouth or through other services that collate discount codes, and users who watch the advertisement may choose to buy the product in the future while forgetting to key in the promo code. These problems with sample selection leads to a threat to internal validity that causes issues to causal inference.

Another way this source of causal impact may be biased is through the omission of confounding variables.
The current regression is Usage_of_Promo_Code = β0 + α*Cost_of_Ads + μi

A possible confounding variable not included could be the popularity of youtube in differing countries (or country/region of viewers). This popularity has an impact on the usage of promo codes as in countries where youtube is more popular, such ads are more likely to be seen, and thus more promo codes are used, resulting in an impact on the usage of promo codes. This popularity also has a correlation with the cost of advertisements as the more technologically advanced or competitive each region is, youtube ads are likely to be more expensive. Thus this poses an omission of confounding variables, leading to omitted variable bias or endogeneity where E[ui | xi] != 0. Thus resulting in biased causal estimates.

Q2)
A more accurate approach to reduce sample selection bias is to store the google account details for each user who views the advertisement in a database. Upon purchasing a fashion product that the promo code is applicable to, the code would automatically apply to those specific user accounts. This reduces the probability of users forgetting to key in the promo code after watching the advertisement and also prevents users from receiving the promo code through word of mouth and keying it into the system. However, such an approach also has its flaws in the possibility of users purchasing the product using a different account or not signing in when watching the youtube ad.

Another approach to resolve this issue as a whole includes the concept of difference-in-differences. The online retailer is able to find a product that is similar to the product they're advertising and has sales similar to the product that has been advertised with the KOL ads. They should observe the sales trend data of both products prior to and after the KOL advertisements have been launched. This allows the similar product that isn't subjected to the KOL ads to act as a form of control group, which presumably absorbs the unobserved differences between treatment and control groups over time. For instance, if any global financial crisis hits that decreases demand for such luxury goods, both products would experience a similar fall, knowing that the sales aren't decreasing due to the ad but through an external agent. Overall, they should use the "control" product to project the supposed change in sales of the product subjected to ads, and compare that projection to the actual change, which would be a better representation of the impact such KOL ads would entail.

Overall this difference-in-differences technique absorbs unobserved factors assuming the changes are the same over time.

Q2 AB Testing in Tikkot's Recommender System

1. Check for Random Assignment

Verification of whether Treatment was randomly assigned, ensuring treatment and control groups are comparable on key variables.

The Key Variables to test include Gender, OS, Registration Date, and Past Activity Time.

First, conduct a chi-square test to verify if the distribution of categorical variables such as Gender and OS (ios_vs_android) is significantly different between treatment and control groups.

The Chi-Square Test examines whether categorical variables (two dimensions) are orthogonal with the values within the rest of the contingency table. Thus, we first

Verify Size of Table:

```
data <- read.csv('C:/Users/isaac/OneDrive/Documents/NUS/School_of_Computing/Y3S1/BT4241 Causal Impact Analysis/Assignments/Assignment1/Tikkot_data.csv')
head(data)
nrow(data)
```

```
[1] 700000
```

Since the table has 700000 instances, it is large enough to verify with the Chi-Square Test.

Chi-Square Test Code:

```
# Chi-Square Test for Gender
table_gender <- table(data$Gender, data$Treatment)
chisq.test(table_gender)

# Chi-Square Test for OS
table_os <- table(data$ios_vs_android, data$Treatment)
chisq.test(table_os)
```

Chi-Square Test Results:

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  table_gender
X-squared = 0.2986, df = 1, p-value = 0.5848


        Pearson's Chi-squared test with Yates' continuity correction

data:  table_os
X-squared = 0.23834, df = 1, p-value = 0.6254
```

The X^2 values are 0.2986 and 0.23834 for Gender and ios_vs_android respectively, this is relatively low and supports the idea that the observed data do not deviate much from what would be expected under the null hypothesis.

The p-values on the other hand, are greater than 0.05, being 0.5848 and 0.6254 for Gender and ios_vs_android respectively, this shows that there is no statistically significant association between Gender and Treatment along with OS and Treatment.

Overall these two results suggest that the treatment and control groups are comparable on the variables tested and support the idea that the treatment was randomly assigned.

Next, I will test the active_time, past_active_time, and registration_date numerical variables through t-tests.

active_time and past_active time code:

```
# Treatment group T-Test
t.test(past_active_time ~ Treatment, data = data)
t.test(active_time ~ Treatment, data = data)
```

Results:

```
        Welch Two Sample t-test

data:  past_active_time by Treatment
t = -1.0158, df = 7e+05, p-value = 0.3097
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.007541110  0.002392576
sample estimates:
mean in group 0 mean in group 1
       10.45575        10.45832


        Welch Two Sample t-test

data:  active_time by Treatment
t = -32.976, df = 699849, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.1209731 -0.1073994
sample estimates:
mean in group 0 mean in group 1
       10.45588        10.57007
```

The fact that the past_active_time p-value is 0.3097, shows that there is no statistical significance int he difference in past_active_time between the treatment and control groups (with the significance level being 5%). This shows that the amount of time users spent on the platform prior to the experiment is similar across both groups, being comparable.

However, the t.test for active_time post-treatment shows a p-value of <2.2e-16 indicates a highly statistically significant difference in active_time between treatment and control groups. This indicates a new recommendation algorithm had a significant impact on the amount of time users spent on the platform after the treatment was applied.

Converting Registration Date to Numerica Variable:

```
# Treatment group T-Test
t.test(past_active_time ~ Treatment, data = data)
t.test(active_time ~ Treatment, data = data)

# Converting Registration Date to Date_Type
data1 <- data
data1$registration_date <- as.Date(data1$registration_date)

# Converting Registration Date to Numeric
data1$days_since_registration <- as.numeric(data1$registration_date - min(data1$registration_date))

# Conduct Test
wilcox.test(days_since_registration ~ Treatment, data = data1)
```

Results:

```
        Wilcoxon rank sum test with continuity correction

data:  days_since_registration by Treatment
W = 6.1206e+10, p-value = 0.5996
alternative hypothesis: true location shift is not equal to 0
```

A high p-value of 0.599 here supports the idea that the treatment and control groups are comparable with respect to the variable tested, and that the observed differences in outcomes are more likely due to the treatment itself rather than pre-existing differences between the groups.

2. Estimate the ATE

Self-Coded t-test:

```
# Grouping data by Treatment
treatment_group <- data %>% filter(Treatment == 1)
control_group <- data %>% filter(Treatment == 0)

# Compute Stats
mean_treatment <- mean(treatment_group$active_time)
mean_control <- mean(control_group$active_time)

var_treatment <- var(treatment_group$active_time)
var_control <- var(control_group$active_time)

n_treatment <- nrow(treatment_group)
n_control <- nrow(control_group)

pooled_se <- sqrt((var_treatment / n_treatment) + (var_control / n_control))
t_statistic <- (mean_treatment - mean_control) / pooled_se
df <- min(n_treatment - 1, n_control - 1)
p_value <- 2 * (1 - pt(abs(t_statistic), df))

# Compute ATE
ATE <- mean_treatment - mean_control

# Print results
cat("ATE:", ATE, "\n")
cat("T-statistic:", t_statistic, "\n")
cat("P-value:", p_value, "\n")
```

Results:

```
ATE: 0.1141862
T-statistic: 32.97586
P-value: 0
```

Self-coded Permutation Test:

```r
# Set number of permutations
n_permutations <- 1000

# Compute observed ATE
observed_ATE <- mean_treatment - mean_control

# Storing ATEs
perm_ATEs <- numeric(n_permutations)

# Permutation loop
for (i in 1:n_permutations) {
  permuted_data <- data
  permuted_data$Treatment <- sample(permuted_data$Treatment) # Shuffle Treatment

  # Compute permuted group means
  perm_treatment_group <- permuted_data %>% filter(Treatment == 1)
  perm_control_group <- permuted_data %>% filter(Treatment == 0)

  perm_mean_treatment <- mean(perm_treatment_group$active_time)
  perm_mean_control <- mean(perm_control_group$active_time)

  # Compute permuted ATE
  perm_ATEs[i] <- perm_mean_treatment - perm_mean_control
}

# Compute p-value as the proportion of permuted ATEs more extreme than observed ATE
p_value_permutation <- mean(abs(perm_ATEs) >= abs(observed_ATE))

# Results
cat("Observed ATE:", observed_ATE, "\n")
cat("Permutation Test P-value:", p_value_permutation, "\n")
```

Results:

```
Observed ATE: 0.1141862
Permutation Test P-value: 0
```

In-built T-Test:

```r
# Built-in t-test for active_time by Treatment
t_test_result <- t.test(active_time ~ Treatment, data = data)

# Results
t_test_result
```

Results:

```
        Welch Two Sample t-test

data:  active_time by Treatment
t = -32.976, df = 699849, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.1209731 -0.1073994
sample estimates:
mean in group 0 mean in group 1
       10.45588        10.57007
```

These results show a p-value of < 2.2e^-16, which is extremely low. This indicates the difference in active_time between treatment and control groups are statistically significant, thus the new recommendation algorithm has a significant effect on the active_time variable.

Regression Analysis Code:

```
# Linear Regression Model
reg_model <- lm(active_time ~ Treatment, data = data)
summary(reg_model)

# Storing Coefficient of Treatment as ATE
ATE_regression <- coef(reg_model)["Treatment"]
cat("ATE from regression:", ATE_regression, "\n")
```

Results:

```
Call:
lm(formula = active_time ~ Treatment, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6068 -0.9781  0.0000  0.9780  7.0474

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.455884   0.002448 4271.93   <2e-16 ***
Treatment    0.114186   0.003463   32.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.449 on 699998 degrees of freedom
Multiple R-squared:  0.001551,  Adjusted R-squared:  0.00155
F-statistic:  1087 on 1 and 699998 DF,  p-value: < 2.2e-16

ATE from regression: 0.1141862
```

As seen in the previous tests, the ATE calculated is at a consistent 0.1141861, which is approximately 0.114 (3s.f.)

Calculations:

```
```{r calcs}
# Calculating the average active_time users spend in the month prior to the experiment
mean(data$past_active_time)
```

 [1] 10.45704
```

Although 0.114 minutes per user may seem to have a small impact at face value, depending on the platform and their user activity, it may be great. As seen in the code above, the average past_active_time prior to this change in algorithm is about 10.5 (3s.f.) in the month prior to the experiment. This leads to an overall 1.09% percent increase. Which although still minimal, is statistically significant and represents a meaningful improvement in user engagement.

3. Role of Gender, Registration Date, and OS

In order to observe the influence of Gender, Registration Date, and Operating System, I’ll build a regression model to first verify their significance (p-value less than 0.05) before viewing their coefficient to determine the impact they make.

Code:

```r
# Converting Registration Date to Numeric
data$registration_date <- as.Date(data$registration_date)
data$days_since_registration <- as.numeric(difftime(data$registration_date, min(data$registration_date), units = "days"))

# Converting categorical variables to factors
data$Gender <- as.factor(data$Gender)
data$ios_vs_android <- as.factor(data$ios_vs_android)

# Linear regression with interaction terms
model_interaction <- lm(active_time ~ Treatment * Gender +
                                      Treatment * days_since_registration +
                                      Treatment * ios_vs_android,
                        data = data)

# Summary of the model
summary(model_interaction)
```

Results:

```
Call:
lm(formula = active_time ~ Treatment * Gender + Treatment * days_since_registration +
    Treatment * ios_vs_android, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2512 -0.9550  0.0005  0.9551  6.6962

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                       10.1411221  0.0189748 534.453  < 2e-16 ***
Treatment                          0.1403193  0.0268332   5.229  1.7e-07 ***
GenderMale                         0.6880802  0.0086317  79.715  < 2e-16 ***
days_since_registration           -0.0002433  0.0001197  -2.033   0.0421 *
ios_vs_androidiOS                  0.0002551  0.0048765   0.052   0.9583
Treatment:GenderMale              -0.1844742  0.0122094 -15.109  < 2e-16 ***
Treatment:days_since_registration -0.0001327  0.0001692  -0.784   0.4330
Treatment:ios_vs_androidiOS        0.1387197  0.0068983  20.109  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.414 on 699992 degrees of freedom
Multiple R-squared:  0.04851,   Adjusted R-squared:  0.0485
F-statistic:  5098 on 7 and 699992 DF,  p-value: < 2.2e-16
```

As seen from the results, based on the newly implemented algorithm, the Gender and Operating System is significant. Both Gender and OS have a p-value of <2e-16, this is far below the standard 5% level of significance, showing that there is a strong indication that the value is not 0.

For Gender, both variables before and after treatment are statistically significant, however, the coefficient reverses in magnitude from 0.688 (3s.f.) with the previous algorithm and -0.184 (3s.f.) with the new algorithm. This shows that being male is associated with 0.688 more minutes of user engagement compared to being female ceteris paribus. Whereas with the new algorithm, being male is associated with 0.184 fewer minutes of user engagement compared to being female ceteris paribus.
Depending on the goal of the company with its target audience, I would recommend the new algorithm if they’ve shifted their target audience more towards women and are willing to forgo more of their male audience. However, as a whole, the magnitude of 0.688 from the previous algorithm is significantly more than the magnitude of the new algorithm being -0.184, and would recommend the previous algorithm assuming ceteris paribus.

For OS, the new algorithm is statistically significant being <2e-16 while the old algorithm isn't, at 0.9583. This shows that the new algorithm affects the OS, while the old algorithm does not. This is supported by the old algorithm’s magnitude being very small, at 0.000243. while the new algorithm is at 0.139 (3s.f.). With R assigning factors alphabetically, this magnitude shows that there is 0.139 minutes more for Android users compared to ios users. Depending on tikkot’s target audience, and possibly the percentage of users.

```
# Calculating percentage of Android Users
counts <- table(data$ios_vs_android)
percentages <- prop.table(counts) * 100
print(percentages)
```

```
 Android      iOS
40.06457 59.93543
```

As per calculations, ios has about 60% of the users in this experiment. Assuming it is reflective of the actual user percentage, I would not recommend the new algorithm in this area as the new algorithm benefits user engagement in android more than ios. Based on proportion, it would lead to about 0.1387197 * 0.4 = 0.0555 (3s.f.) increase in user engagement as a whole which isn't as significant as other areas.

In terms of Registration Date, it was initially significant for the previous algorithm at 0.0421 p-value but not significant with the new algorithm at 0.433. The magnitude for both is relatively small at -0.000255 in the previous algorithm and -0.000133 with the new algorithm. Overall this is too small of a change for me to make recommendations to cause significant change.

Overall, based on these calculations on the regression analysis, I would not recommend switching to the new algorithm as the new algorithm leads to the Gender, OS and Registration Date having much less significant impact on the active_time, forgoing the significant impact Gender brought to the previous algorithm, assuming they do not intend to switch target audiences ceteris paribus. However, if they're looking to expand more into targeting women, or looking at pure user_activity, they may consider using the new algorithm as the ATE increases. With that said, they should run more experiments with the new algorithm to find variables that lead to a greater impact on the active_time.

Q3 Demand Estimation and Pricing for Grad

1. OLS Regression Discussion

Variables I'll include in my OLS Regression:

- q, the quantity or number of rides as the dependent variable
- p, the post-discount price as the main independent variable
- weekend, as the binary variable indicates if its the weekend
    - Weekend variable has an impact on the quantity of rides, as weekends provide greater accessibility to users as they're free from work and are able to travel (thus increasing quantity)
    - Weekend variable is correlated to the post-discount price as pre-discount prices are bound to increase during the weekends as there's greater demand for rides as people are free from work
- Location, as the categorical variable to determine the region of the ride
    - Location has an impact on the quantity of rides as more popular places like MBS or Clarke Quay is bound to have more commuters and thus more demand
    - Location also has a correlation to post-discount price as the pre-discount price is bound to increase at specific more popular locations.
- ε as the error term

OLS Regression: $q = \beta0 + \beta1*p + \beta2*weekend + \beta3*location + \varepsilon$

Take note that my actual code includes categorical_weekend and categorical_location to be able to run the regression.

Thus the results include:$q = \beta0 + \beta1*p + \beta2*categorical\_weekend1 + \beta3*categorical\_locationMBS + \beta4*categorical\_locationOrchard + \varepsilon$

Code:

```
data <- read.csv('C:/Users/isaac/OneDrive/Documents/NUS/School_of_Computing/Y3S1/BT4241 Causal Impact Analysis/Assignments/Assignment1/grad.csv')
head(data)

data$categorical_weekend <- factor(data$weekend, ordered = FALSE)
data$categorical_location <- factor(data$Location, ordered = FALSE)
regression <- lm(q ~ p + categorical_weekend + categorical_location, data = data)
summary(regression)
```

Results:

```
Call:
lm(formula = q ~ p + categorical_weekend + categorical_location,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8033 -0.6670  0.0002  0.6779  3.8493

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              21.242190   0.057178  371.51   <2e-16 ***
p                        -0.405678   0.006733  -60.25   <2e-16 ***
categorical_weekend1      0.527003   0.039177   13.45   <2e-16 ***
categorical_locationMBS   3.001194   0.024256  123.73   <2e-16 ***
categorical_locationOrchard  0.009207   0.024233    0.38    0.704
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9892 on 9995 degrees of freedom
Multiple R-squared:  0.7471,    Adjusted R-squared:  0.747
F-statistic:  7383 on 4 and 9995 DF,  p-value: < 2.2e-16
```

The estimates include the intercept β0 and the coefficient of the p variable, β1, coefficient of categorical_weekend1, β2, coefficient of categorical_locationMBS, β3 coefficient of categorical_locationOrchard, β4. All are statistically significant at <2e-16, which is much smaller than 5% accepted level of significance, other than categorical_locationOrchard at 0.704, this indicates that the location of Orchard is not significant in determining the quantity. The intercept is positive, at 21.2 (3s.f.), which indicates even without any discounts, there is a baseline quantity of rides that will still be present. The price estimate of -0.406 (3s.f.) indicates that for every unit increase in discounts is associated with a 0.406 decrease in the quantity of rides. This makes sense as lower the price, there tends to be greater quantity demanded for the good, and thus an increase in unit leads to less quantity of rides.

Categorical_weekend1 has an estimate of 0.527, which indicates that weekends are associated with a 0.527 increase in quantity of rides, ceteris paribus.
Categorical_locationMBS having an estimate of 3.00 (3s.f.) indicates that if the region of the ride is in MBS, it is associated with a 3.00 increase in the quantity of rides, ceteris paribus.

2. Endogeneity Explanation

The estimates from the OLS regression might be problematic as the main independent variable, p, may consider real-time demand conditions that are not recorded in the data. This results in E[ε | p] != 0, having omitted variable bias such as weather which has an impact on the number of rides as bad weather such as rain would increase quantity, along with bad weather having a correlation to price as more demand increases, resulting in higher prices. This omitted variable bias, or endogeneity goes against the causal assumption of E[ε | p] != 0, leading to a threat to internal validity. In order to combat this, we would normally include such confounding variables in the regression to control them, however, real-time demand conditions include possibly unobservable variables which would make them hard to capture.

3. Instrumental Variable Approach

One Instrument Variable(IV) I would propose is the Discount Variable.
In terms of the conditions for a good IV, instrument relevance and instrument exogeneity are required.
For Instrument Exogeneity, discounts are “totally randomized”. This randomization makes it so that there is no correlation of discounts to any other variable, where this lack of correlation leads to no omission of confounding variables as confounding variables are those that have an impact on q and are correlated to the discounts.
For Instrument Relevance, the Discount Variable is correlated with the quantity of rides.

Running IV Regression:

```
library(AER)

# Regress p on the instrument and other exogenous variables
first_stage <- lm(p ~ discounts + weekend + Location, data = data)
summary(first_stage)

# Use the predicted values from the first stage in the main regression
data$p_hat <- predict(first_stage)

# IV regression
iv_regression <- ivreg(q ~ p_hat + weekend + Location | discounts + weekend + Location, data = data)
summary(iv_regression)
```

Results (First Stage):

```
Call:
lm(formula = p ~ discounts + weekend + Location, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0902 -0.8181 -0.0064  0.8269  4.9229

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.99723    0.03878 257.824   <2e-16 ***
discounts        1.00382    0.01494  67.204   <2e-16 ***
weekend          5.01620    0.02439 205.642   <2e-16 ***
LocationMBS      0.01156    0.02991   0.387    0.699
LocationOrchard -0.03858    0.02988  -1.291    0.197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.22 on 9995 degrees of freedom
Multiple R-squared:  0.8244,    Adjusted R-squared:  0.8243
F-statistic: 1.173e+04 on 4 and 9995 DF,  p-value: < 2.2e-16
```

Results (IV Regression):

```
Call:
ivreg(formula = q ~ p_hat + weekend + Location | discounts +
    weekend + Location, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-3.922823 -0.720680  0.007445  0.731123  3.930599

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     22.030902   0.107034 205.830   <2e-16 ***
p_hat           -0.504645   0.013156 -38.358   <2e-16 ***
weekend          1.024033   0.069503  14.734   <2e-16 ***
LocationMBS      3.004247   0.026443 113.611   <2e-16 ***
LocationOrchard  0.005901   0.026419   0.223    0.823
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.078 on 9995 degrees of freedom
Multiple R-Squared: 0.6995,     Adjusted R-squared: 0.6994
Wald test:  5818 on 4 and 9995 DF,  p-value: < 2.2e-16
```

The p-value for discounts and p-hat are <2e-16, making it highly statistically significant. This high statistical significance implies that discounts are correlated with the post-discount price, thus having instrument relevance.

With both Instrument Relevance and Instrument Exogeneity, Price is able to impact the Quantity of rides solely through the Post-discount Price without any other variables and thus would work as a good Instrument Variable.

4. Optimal Pricing Analysis

To verify if prices are based on time and location, I will first conduct a regression analysis to verify if they're correlated:
Code:

```
#Verify Correlation betweeen Price against Time and Location
regression_price <- lm(p ~ weekend + Location, data = data)
summary(regression_price)
```

Results:

```
Call:
lm(formula = p ~ weekend + Location, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0000 -0.9954 -0.0077  0.9838  5.0233

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.96941    0.02934 271.604   <2e-16 ***
weekend          5.02215    0.02939 170.879   <2e-16 ***
LocationMBS      0.03085    0.03603   0.856    0.392
LocationOrchard -0.03341    0.03600  -0.928    0.353
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.469 on 9996 degrees of freedom
Multiple R-squared:  0.745,     Adjusted R-squared:  0.7449
F-statistic:  9735 on 3 and 9996 DF,  p-value: < 2.2e-16
```

The fact that the weekend variable has a p-value of <2e-16 shows that it is statistically significant and the null hypothesis that the coefficient of the weekend variable being 0 is false. This implies that there is a relatively strong correlation between the weekend variable and price, not to mention the relatively large coefficient of 5.02 (3s.f.)

The Location variable on the other hand has neither LocationMBS nor LocationOrchard being statistically significant, being 0.392 and 0.353 respectively. Since they are not statistically significant, the pricing team likely set prices only based on time (weekend) and not location.

To Maximize $\pi(p) = (p - mc) * q(p, weekend, location)$
In this instance, p, weekend, and location are β1, β2, and β3 respectively, with mc == 5
Thus, $\pi(p) = (p - mc) * (\beta0 + \beta1 * p + \beta2 * weekend + \beta3 * Location)$

In order to maximize this function, I will find its derivative and equate it to 0.
$d\pi(p)/dp = (2 * \beta1 * p + \beta0 + \beta2 * weekend + \beta3 * Location) - mc * \beta1 = 0$

Thus,
$p = (mc * \beta1 - \beta0 - \beta2 * weekend - \beta3 * Location) / (2\beta1)$

Implementing Optimal Price Formula:

```
data <- read.csv('C:/Users/isaac/OneDrive/Documents/NUS/School_of_Computing/Y3S1/BT4241 Causal Impact Analysis/Assignments/Assignment1/grad.csv')
head(data)

data$categorical_weekend <- factor(data$weekend, ordered = FALSE)
data$categorical_location <- factor(data$Location, ordered = FALSE)
regression <- lm(q ~ p + categorical_weekend + categorical_location, data = data)
summary(regression)

iv_reg <- lm(p ~ discounts, data = data)
summary(iv_reg)

#Verify Correlation betweeen Price against Time and Location
regression_price <- lm(p ~ weekend + Location, data = data)
summary(regression_price)

# Marginal Cost Stated is 5
mc <- 5

# Performing Regression to Attain Demand Function Coefficients
demand_model <- lm(q ~ p + categorical_weekend + categorical_location, data = data)
summary(demand_model)

# Extract coefficients
coeffs <- coefficients(demand_model)
beta_0 <- coeffs[1]
beta_1 <- coeffs[2]
beta_2 <- coeffs[3]
# Assume locations have three levels and coefficients for them are:
beta_3_orchard <- coeffs["categorical_locationOrchard"]
beta_3_mbs <- coeffs["categorical_locationMBS"]

# Defining Optimal Price Function
optimal_price <- function(beta_0, beta_1, beta_2, beta_3, mc) {
  return((mc * beta_1 - beta_0 - beta_2 - beta_3) / (2 * beta_1))
}

# Calculating Optimal Price
optimal_price_weekend_orchard <- optimal_price(beta_0, beta_1, beta_2, beta_3_orchard, mc)
optimal_price_weekend_mbs <- optimal_price(beta_0, beta_1, beta_2, beta_3_mbs, mc)
optimal_price_weekend_clarke_quay <- optimal_price(beta_0, beta_1, beta_2, 0, mc)
optimal_price_weekday_orchard <- optimal_price(beta_0, beta_1, 0, beta_3_orchard, mc)
optimal_price_weekday_mbs <- optimal_price(beta_0, beta_1, 0, beta_3_mbs, mc)
optimal_price_weekday_clarke_quay <- optimal_price(beta_0, beta_1, 0, 0, mc)
```

^since clarke_quay is the reference category, there is no need to include a beta_3

Results:

```
Optimal Price on Weekend (Orchard): 29.34199
Optimal Price on Weekend (MBS): 33.02963
Optimal Price on Weekend (Clarke Quay): 29.33064
Optimal Price on Weekday (Orchard): 28.69245
Optimal Price on Weekday (MBS): 32.38009
Optimal Price on Weekday (Clarke Quay): 28.68111
```

When comparing between the Optimal and Current Prices:

```
# Load the data
data <- read.csv('C:/Users/isaac/OneDrive/Documents/NUS/School_of_Computing/Y3S1/BT4241 Causal Impact Analysis/Assignments/Assignment1/grad.csv')

# Converting Categorical Variables to Factors
data$weekend <- factor(data$weekend, ordered = FALSE)
data$Location <- factor(data$Location, ordered = FALSE)

# Computing Average Price for Weekend in Orchard
avg_price_weekend_orchard <- data %>%
  filter(weekend == 1, Location == "Orchard") %>%
  summarize(avg_price = mean(p, na.rm = TRUE))

# Computing Average Price for Weekend in MBS
avg_price_weekend_mbs <- data %>%
  filter(weekend == 1, Location == "MBS") %>%
  summarize(avg_price = mean(p, na.rm = TRUE))

# Computing Average Price for Weekend in Clarke Quay
avg_price_weekend_clarke_quay <- data %>%
  filter(weekend == 1, Location == "Clarke Quay") %>%
  summarize(avg_price = mean(p, na.rm = TRUE))

# Computing Average Price for Weekday in Orchard
avg_price_weekday_orchard <- data %>%
  filter(weekend == 0, Location == "Orchard") %>%
  summarize(avg_price = mean(p, na.rm = TRUE))

# Computing Average Price for Weekday in MBS
avg_price_weekday_mbs <- data %>%
  filter(weekend == 0, Location == "MBS") %>%
  summarize(avg_price = mean(p, na.rm = TRUE))

# Computing Average Price for Weekday in Clarke Quay
avg_price_weekday_clarke_quay <- data %>%
  filter(weekend == 0, Location == "Clarke Quay") %>%
  summarize(avg_price = mean(p, na.rm = TRUE))

# Compiling Results
average_prices <- data.frame(
  Condition = c("Weekend in Orchard", "Weekend in MBS", "Weekend in Clarke Quay",
                "Weekday in Orchard", "Weekday in MBS", "Weekday in Clarke Quay"),
  Average_Price = c(avg_price_weekend_orchard$avg_price, avg_price_weekend_mbs$avg_price, avg_price_weekend_clarke_quay$avg_price,
                    avg_price_weekday_orchard$avg_price, avg_price_weekday_mbs$avg_price, avg_price_weekday_clarke_quay$avg_price)
)
```

Results:

Condition <chr>	Average_Price <dbl>
Weekend in Orchard	12.931133
Weekend in MBS	13.058808
Weekend in Clarke Quay	12.982249
Weekday in Orchard	7.963612
Weekday in MBS	7.963517
Weekday in Clarke Quay	7.978479

The Optimal Prices listed are significantly larger than the current prices.
Certain misalignments include the optimal weekend prices not being too different from the optimal weekday prices, which could be due to the β2 having a relatively small magnitude of 0.527 compared to certain location variables such as categorical_locationMBS which has a coefficient of 3.00.
Whereas the current price for weekends are all significantly greater than the weekdays.

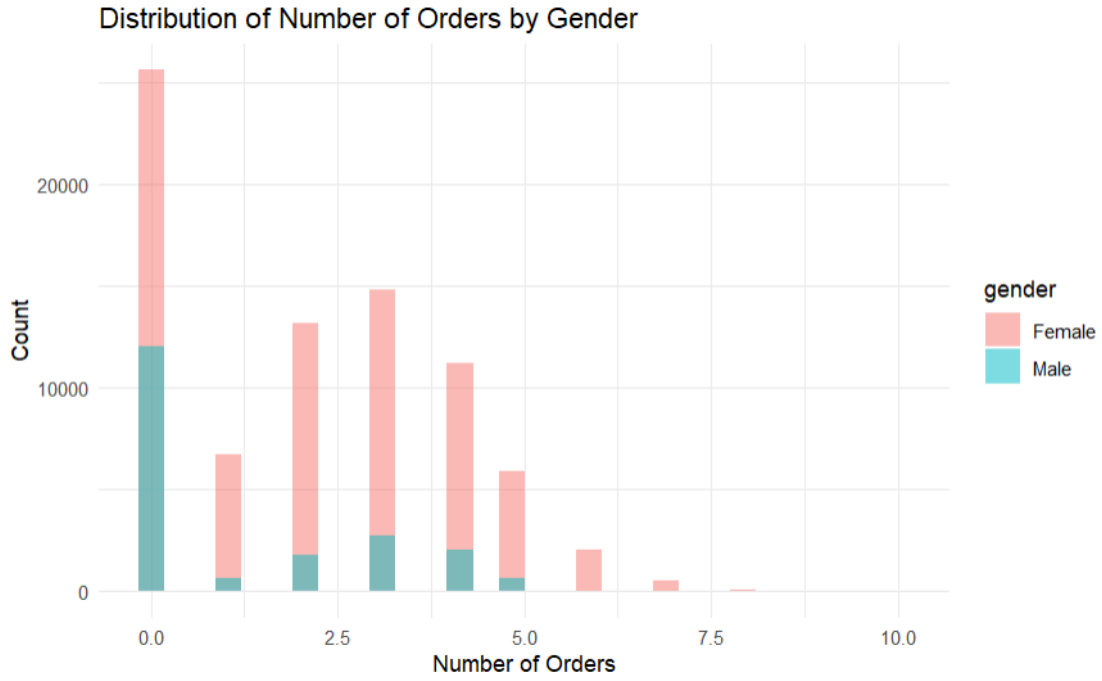
Q4) Experiment Design in Ladaza

1. Visualization of Distribution

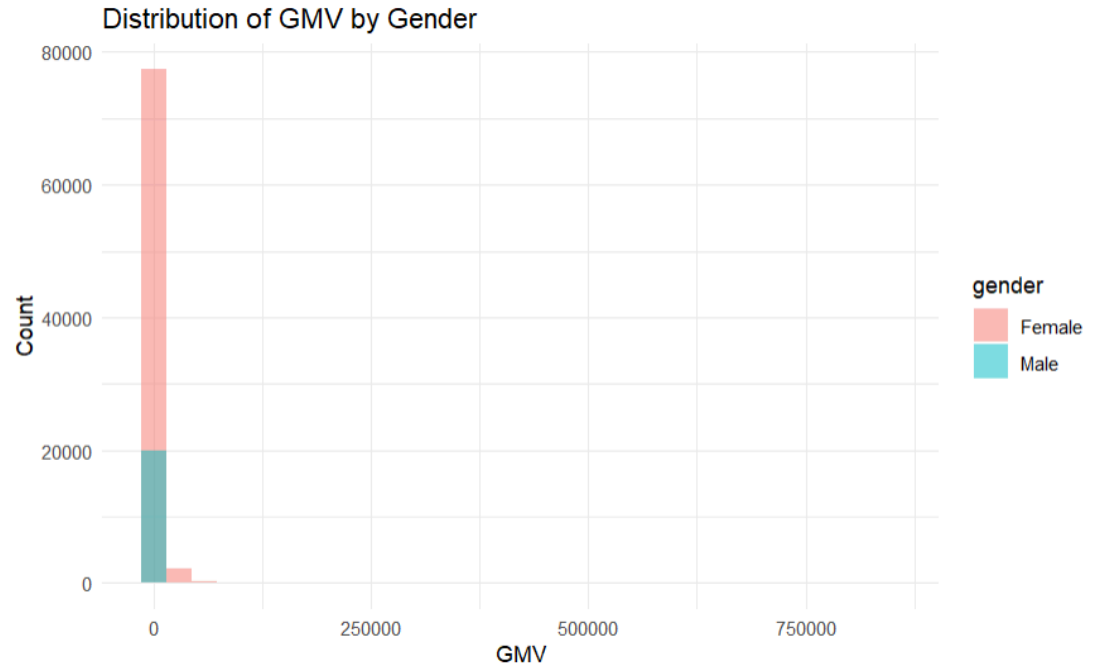
Code:

```
library(ggplot2)

data <- read.csv('C:/Users/isaac/OneDrive/Documents/NUS/School_of_Computing/Y3S1/BT4241 Causal Impact
Analysis/Assignments/Assignment1/ladaza.csv')
head(data)

# Plotting the distribution of the number of orders by gender
ggplot(data, aes(x = number_of_orders, fill = gender)) +
  geom_histogram(alpha = 0.5, position = "identity", bins = 30) +
  labs(title = "Distribution of Number of Orders by Gender",
       x = "Number of Orders",
       y = "Count") +
  theme_minimal()

# Plotting the distribution of GMV by gender
ggplot(data, aes(x = gmv, fill = gender)) +
  geom_histogram(alpha = 0.5, position = "identity", bins = 30) +
  labs(title = "Distribution of GMV by Gender",
       x = "GMV",
       y = "Count") +
  theme_minimal()
```

Distribution of Number of Orders by Gender Results:



Distribution of GMV by Gender Results:



2. Sample Size Calculation

Computing Sample Size Code:

```r
# Define parameters
effect_size <- 0.007  # 0.7% effect
power <- 0.80         # 80% power
alpha <- 0.05         # 5% significance level

# Compute the required sample size
sample_size <- pwr.p.test(h = effect_size, sig.level = alpha, power = power, alternative = "two.sided")

# Print the required sample size
print(sample_size)
```

Results:

```
   proportion power calculation for binomial distribution (arcsine transformation)

           h = 0.007
           n = 160180.8
   sig.level = 0.05
       power = 0.8
 alternative = two.sided
```

Thus the treatment sample group size would be around 160180 people.

To separate the Males and Females to attain Control Group Size, I'll first calculate the proportions:

Code:

```r
proportions <- prop.table(table(data$gender))
print(proportions)
```

Results:

```
    Male  Female
 0.19933 0.80067
```

Thus, the treatment sample group size for males would be 160180*0.19933 = 31929 (rounded up), with the treatment sample group size for females being 160180 - 31929 = 128251 based on proportion.

Q5) A Brilliant Idea

1. Run AA Tests

Pre-given Code Results (Non-static):

```
$p_value
[1] 0.213452

$total_users
[1] 500

$contingency_table
                                 control_sucess control_total - control_success
treatment_success                           137                             119
treatment_total - treatment_success         117                             127
```

Edited Code to perform a series of AA tests to attain false positive rate:

```r
# Set parameters
p <- 0.5   # Probability of assigning to treatment
max_N <- 500  # Maximum number of users
alpha <- 0.05  # Significance level
num_simulations <- 100  # Number of simulations to run

# Run multiple AA tests
false_positives <- sum(replicate(num_simulations, simulate_ab_test(p, max_N, alpha)))

# Calculate the false positive rate
false_positive_rate <- false_positives / num_simulations

# Print the false positive rate
cat("False Positive Rate:", false_positive_rate, "\n")
```

False Positive Rate Results: 0.68

2. Present Results

Table of AA Tests with varying Max_N:

Max_N	500	1000	1500	2000
Total_Users	500	1000	1500	2000
Size of Control Group	256	508	746	1028
P_Value	0.213452	0.5643288	0.2756984	0.04974907

As my Max_N value increases, the size of the control group increases as well. This is due to the condition of assessing identical control groups, about 50% of the Total_Users will be assigned to the control group and the remaining assigned to the treatment group.

The p-value on the other hand has no specific trend, increasing from 0.213452 to 0.5643288 before lowering to 0.2756984 and 0.04974907. This is expected as the p-value in this simulation represents a false positive since both control and treatment groups are identical. This lack of trend shows that the false positiveness is inherently random when there is no true effect.

3. Critique the Method

One main issue with this experiment is that false positive rates are inherently random. The current method stops the simulation the moment a p-value of less than 0.05 is attained, which means it can occur at any time. As the simulation has varying total_users, thus varying sample size, the possibility of the p-value being less than 0.05 and stopping the experiment is random. If the simulation is run about 100 times with the proportion of p-values falling less than 0.05 calculated, this would provide a more accurate method, as the law of large numbers comes into play and the randomness of the p-value being less than 0.05 resulting in a false positive having a smaller effect. Stopping it immediately could simply be a result of luck and lead to misleading conclusions based on insufficient data (as many times, the experiment stops with a single or low double-digit sample size.