How coupon redemption affect business?

RangLi

2023-08-11

Introduction

Coupons have been using for business as a tool for promotion and gain sales for several decades. With the development of technology, coupon delivering methods have been changing, from mailing paper coupons to sending mobile coupons, as well as coupon codes sent by e-commerce. Coupons could active previous customers, also brings in new customers. Promotions by retail companies attracts consumers to try new products or stock items that is on sale. It's estimated that 60% of consumers are likely to try a new product as a result of a coupon, while 46% said they'd change their planned purchase(Epstein, 2022). In the meantime, business interests in how coupons have benefit there sales, as well as how to minimize their investment on the promotion campaign to gain the most profit.

As one of the most practical way to increase the revenue and enhance the profit margin, essentially each merchant is keen to understand how to boost their business by sending coupons more effectively. Analyzing the previous coupon redemption data systematically will be a very intersting and important approach.

Research questions

1. Correspondingly how could we clean the dataset to keep good data?
2. What data fields are most relevant for this study?
3. How should we categorize the data set, how many ways we could categorize it?
4. Is there any certain distribution pattern for the dataset itself?
5. Did we find any statistical correlation between two groups?
6. What is the return on investment from coupons? and what kind of return that is?
7. What group of consumers use coupons more?

Approach

First, I'm going to investigate what category has the most redemption of coupons base on the dataset.

Secondly, I'll further investigate customer demographics dataset to summarize the groups of customers that had the most coupon redemption.

Thirdly, I'll make some analysis based on the customer transaction data set on how well coupons are redeemed and make the comparison of the two groups on which group has contributed more to business sales.

The data set we are working on first are "customer_demographics" and "customer_transaction_data".

The customer_demographics data has 6 columns/dimensions, there is no missing values in column "age_range", "rented", "family_size" and "income_bracket"; however, there are lots of missing values on column "marital_status", "no_of_children".

Figure 1: coupon-main

Each column/dimension, in real life perspective, may have an actual impact on the output, we will perform a model training with the 4 columns with complete input to start with, on the other hand, we could also get rid of entries with empty marital status, in that case, we will be able to train with 5 columns.

```
##   customer_id age_range marital_status rented family_size no_of_children
## 1           1       70+        Married      0           2
## 2           6     46-55        Married      0           2
## 3           7     26-35                     0           3              1
## 4           8     26-35                     0           4              2
## 5          10     46-55         Single      0           1
## 6          11       70+         Single      0           2
##   income_bracket
## 1              4
## 2              5
## 3              3
## 4              6
## 5              5
## 6              1
```

```
##    customer_id age_range marital_status rented family_size no_of_children
## 17          31     36-45         Single      0          5+             3+
## 18          33     46-55        Married      0          5+             3+
## 23          40     56-70        Married      0           4              2
## 25          42     26-35        Married      0           4              2
## 26          45     46-55        Married      0          5+             3+
## 29          52     36-45        Married      0          5+             3+
##    income_bracket
## 17              2
## 18              9
## 23              7
## 25              9
## 26              1
## 29              7
```

I've replaced the empty values to NA, and used na.omit to omit all the rows contains NA, so that we only keep the rows that have completed data. I assigned the new data set to "demo_df_1".

From the cleaned data, I would like to do some grouping, to group them by ages. I would like to set ages between 18-25 as group1, ages 26-35 as group2, 36-45 as group3, 46-55 as group4, 56-70 as group 5, and ages 70 and older as group 6. By grouping them into different age groups, we could make some analysis see which group has the most customer.

```
group1 <- demo_df_1 %>% filter(age_range=="18-25")
group2 <- demo_df_1 %>% filter(age_range=="26-35")
group3 <- demo_df_1 %>% filter(age_range=="36-45")
group4 <- demo_df_1 %>% filter(age_range=="46-55")
group5 <- demo_df_1 %>% filter(age_range=="56-70")
group6 <- demo_df_1 %>% filter(age_range=="70+")
summary(group1)
```

```
##   customer_id      age_range         marital_status         rented
##  Min.   : 110.0   Length:9           Length:9           Min.   :0.0000
##  1st Qu.: 775.0   Class :character   Class :character   1st Qu.:0.0000
```

```
##  Median : 894.0   Mode  :character   Mode  :character   Median :1.0000
##  Mean   : 802.7                                        Mean   :0.5556
##  3rd Qu.:1023.0                                        3rd Qu.:1.0000
##  Max.   :1131.0                                        Max.   :1.0000
##  family_size        no_of_children     income_bracket
##  Length:9           Length:9           Min.   :1.000
##  Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Median :4.000
##                                        Mean   :3.222
##                                        3rd Qu.:5.000
##                                        Max.   :6.000
```

```
summary(group2)
```

```
##   customer_id      age_range         marital_status         rented
##  Min.   :  42.0   Length:36          Length:36          Min.   :0.0000
##  1st Qu.: 440.0   Class :character   Class :character   1st Qu.:0.0000
##  Median : 894.5   Mode  :character   Mode  :character   Median :0.0000
##  Mean   : 815.4                                        Mean   :0.1667
##  3rd Qu.:1194.2                                        3rd Qu.:0.0000
##  Max.   :1520.0                                        Max.   :1.0000
##  family_size        no_of_children     income_bracket
##  Length:36          Length:36          Min.   : 1.000
##  Class :character   Class :character   1st Qu.: 4.000
##  Mode  :character   Mode  :character   Median : 4.500
##                                        Mean   : 4.806
##                                        3rd Qu.: 6.000
##                                        Max.   :11.000
```

```
summary(group3)
```

```
##   customer_id      age_range         marital_status         rented
##  Min.   :  31.0   Length:65          Length:65          Min.   :0.00000
##  1st Qu.: 327.0   Class :character   Class :character   1st Qu.:0.00000
##  Median : 533.0   Mode  :character   Mode  :character   Median :0.00000
##  Mean   : 694.1                                        Mean   :0.07692
##  3rd Qu.:1202.0                                        3rd Qu.:0.00000
##  Max.   :1558.0                                        Max.   :1.00000
##  family_size        no_of_children     income_bracket
##  Length:65          Length:65          Min.   : 1.000
##  Class :character   Class :character   1st Qu.: 4.000
##  Mode  :character   Mode  :character   Median : 5.000
##                                        Mean   : 5.554
##                                        3rd Qu.: 7.000
##                                        Max.   :12.000
```

```
summary(group4)
```

```
##   customer_id      age_range         marital_status         rented
##  Min.   :  33.0   Length:46          Length:46          Min.   :0.00000
##  1st Qu.: 541.8   Class :character   Class :character   1st Qu.:0.00000
##  Median : 754.0   Mode  :character   Mode  :character   Median :0.00000
```

```
##  Mean   : 802.6                                          Mean   :0.04348
##  3rd Qu.:1082.8                                          3rd Qu.:0.00000
##  Max.   :1578.0                                          Max.   :1.00000
##  family_size        no_of_children     income_bracket
##  Length:46          Length:46          Min.   :1.000
##  Class :character   Class :character   1st Qu.:4.000
##  Mode  :character   Mode  :character   Median :5.000
##                                        Mean   :4.957
##                                        3rd Qu.:6.000
##                                        Max.   :9.000
```

```
summary(group5)
```

```
##   customer_id      age_range         marital_status         rented
##  Min.   :  40.0   Length:7           Length:7           Min.   :0
##  1st Qu.: 346.5   Class :character   Class :character   1st Qu.:0
##  Median : 474.0   Mode  :character   Mode  :character   Median :0
##  Mean   : 653.1                                        Mean   :0
##  3rd Qu.: 995.5                                        3rd Qu.:0
##  Max.   :1374.0                                        Max.   :0
##  family_size        no_of_children     income_bracket
##  Length:7           Length:7           Min.   :4.000
##  Class :character   Class :character   1st Qu.:4.000
##  Mode  :character   Mode  :character   Median :5.000
##                                        Mean   :5.286
##                                        3rd Qu.:6.500
##                                        Max.   :7.000
```

```
summary(group6)
```

```
##   customer_id     age_range         marital_status         rented
##  Min.   :402.0   Length:2           Length:2           Min.   :0
##  1st Qu.:437.8   Class :character   Class :character   1st Qu.:0
##  Median :473.5   Mode  :character   Mode  :character   Median :0
##  Mean   :473.5                                         Mean   :0
##  3rd Qu.:509.2                                         3rd Qu.:0
##  Max.   :545.0                                         Max.   :0
##  family_size        no_of_children     income_bracket
##  Length:2           Length:2           Min.   :4.00
##  Class :character   Class :character   1st Qu.:4.75
##  Mode  :character   Mode  :character   Median :5.50
##                                        Mean   :5.50
##                                        3rd Qu.:6.25
##                                        Max.   :7.00
```

By looking into each groups, we could see group 4 is the largest customer group, which the age range between 46-55. Meanwhile, group3 and group6 has the largest mean in income_bracket of 5.5, as 10 is the highest income index. We could assume that ages 46-55 tends to shop more, so that we could send more coupons to them to attract those groups of customers to shopping.

"customer_transaction_data" is a complete data set. I would like to work on this data set, using "item id", "selling price", "coupon discount" to see how many items sold with coupon redemption.

```r
transaction <- read.csv("customer_transaction_data.csv")
head(transaction)
```

```
##         date customer_id item_id quantity selling_price other_discount
## 1 2012-01-02        1501   26830        1         35.26         -10.69
## 2 2012-01-02        1501   54253        1         53.43         -13.89
## 3 2012-01-02        1501   31962        1        106.50         -14.25
## 4 2012-01-02        1501   33647        1         67.32           0.00
## 5 2012-01-02        1501   48199        1         71.24         -28.14
## 6 2012-01-02        1501   57397        1         71.24         -28.14
##   coupon_discount
## 1               0
## 2               0
## 3               0
## 4               0
## 5               0
## 6               0
```

```r
transaction_coupon <- transaction %>% select("item_id", "selling_price", "coupon_discount")

transaction_coupon1 <- transaction %>% select("item_id", "selling_price", "coupon_discount") %>% filter
head(transaction_coupon1)
```

```
##   item_id selling_price coupon_discount
## 1    5525        106.50          -35.62
## 2    8145         39.18          -14.25
## 3   16381         48.80          -35.62
## 4   17861         75.51          -26.71
## 5   19583        124.67          -35.62
## 6   20697         92.26          -35.62
```

```r
transaction_coupon2 <- transaction %>% select("item_id", "selling_price", "coupon_discount") %>% filter
head(transaction_coupon2)
```

```
##   item_id selling_price coupon_discount
## 1   26830         35.26               0
## 2   54253         53.43               0
## 3   31962        106.50               0
## 4   33647         67.32               0
## 5   48199         71.24               0
## 6   57397         71.24               0
```

```r
nrow(transaction_coupon1)
```

```
## [1] 21286
```

```r
nrow(transaction_coupon2)
```

```
## [1] 1303280
```

```
nrow(transaction)
```

```
## [1] 1324566
```

After some data transformation, by splitting the customer transaction data into 2 groups, transaction_coupon1 is the group that have used coupons when made a purchase, while transaction_coupon2 is the group that used zero coupons with purchases. By pulling out the datas, we could see that transaction_coupon1 contains 21286 items, means there are 21286 items was sold with a coupon redemption. Meanwhile there are 1303280 purchases was made by a coupon redemption. The total transaction data has 1324566 items, overall there is only 1.6% of purchases are made with a coupon redemption, which is so low compared to the total purchases.

For now, I'm think to merge the "customer_demographics data" with the "customer transaction_data" to better looking at the purchasing power of each group and how well each group redeemed coupons so that business could refer to when they need to send coupons to targeted customer groups. However, we haven't covered that part to merge two data sets and merge them by dividing them using a specific condition.

I've found another data set that could help dig further on this topic.

This is a data set used to make prediction on coupon redemption by a business to compare both online and offline coupon usage. The data I choose is the offline data set,which contains variables "User_id", "Merchant_id", "Coupon_id", "Discount_rate", "Distance", "Date_received" and "Date".

```
offline <- read.csv("offline_train.csv")
head(offline)
```

```
##   User_id Merchant_id Coupon_id Discount_rate Distance Date_received     Date
## 1 1439408        2632      null          null        0          null 20160217
## 2 1439408        4663     11002        150:20        1      20160528     null
## 3 1439408        2632      8591          20:1        0      20160217     null
## 4 1439408        2632      1078          20:1        0      20160319     null
## 5 1439408        2632      8591          20:1        0      20160613     null
## 6 1439408        2632      null          null        0          null 20160516
```

First, I would like to make some changes about the data, I noticed that for the discount rate, the original data has the discount rate marked as, 150:20, which means $20 off $150, while 20:1 means take $1off for evey $20 spent. The way it indicates the discount makes it hard to do analysis for the following steps, so I would like to transform the discount rate into the simple discount rate as 1 indicates 100% of original price, and 0.5 indicates 50% of original price, while 0.95 indicates 5% off. After some transformation, I have the data showing below.

```
offline_df <- read.csv("offline_train2.csv")
head(offline_df)
```

```
##   X User_id Merchant_id Coupon_id Discount_rate Distance Date_received     Date
## 1 0 1439408        2632         0          0.00        0            NA 20160217
## 2 1 1439408        4663     11002          0.87        1      20160528       NA
## 3 2 1439408        2632      8591          0.95        0      20160217       NA
## 4 3 1439408        2632      1078          0.95        0      20160319       NA
## 5 4 1439408        2632      8591          0.95        0      20160613       NA
## 6 5 1439408        2632         0          0.00        0            NA 20160516
```

For this data set, I would like to investigate what is correlated with coupon redemption, I assume both discount rate and distance customers live to the store have a correlation with coupon redemption. I would slice the data into several chunks so that we could make analysis on different conditions.

First of all, I would like to remove all the rows that coupon not received by customers, we only wants to look at the cases that customer redeemed the coupon or not as they received the coupons.

```
offline_df <- offline_df[complete.cases(offline_df[ , ('Date_received')]), ]
head(offline_df)
```

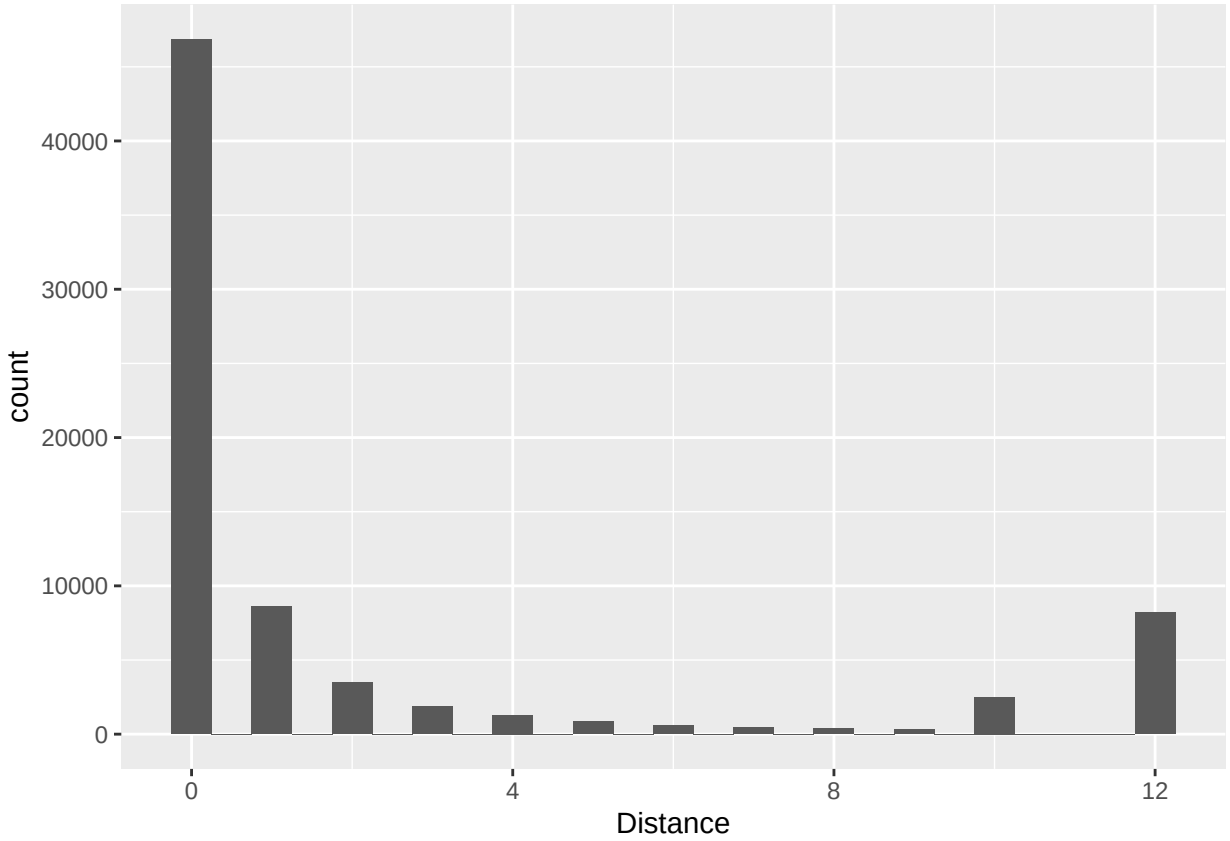
```
##   X User_id Merchant_id Coupon_id Discount_rate Distance Date_received     Date
## 2 1 1439408        4663     11002          0.87        1      20160528       NA
## 3 2 1439408        2632      8591          0.95        0      20160217       NA
## 4 3 1439408        2632      1078          0.95        0      20160319       NA
## 5 4 1439408        2632      8591          0.95        0      20160613       NA
## 7 6 1439408        2632      8591          0.95        0      20160516 20160613
## 8 7 1832624        3381      7610          0.90        0      20160429       NA
```

Then, I would choose the "Date received", "Date"(redeemed) and "distance" as a data frame to see if there is a correlation between those variables. If variable "Date" is not "null" values, that indicates that the coupon is received and redeemed by customers.

```
distance_df <- offline_df %>% select("Distance", "Date_received", "Date") %>% filter(Date != "NA")
head(distance_df)
```

```
##   Distance Date_received     Date
## 1        0      20160516 20160613
## 2        0      20160515 20160521
## 3        0      20160321 20160329
## 4        0      20160523 20160605
## 5        0      20160127 20160221
## 6        0      20160207 20160218
```

```
distance_plot <- ggplot(distance_df, aes(x = Distance)) + geom_histogram(binwidth =0.5)
distance_plot
```



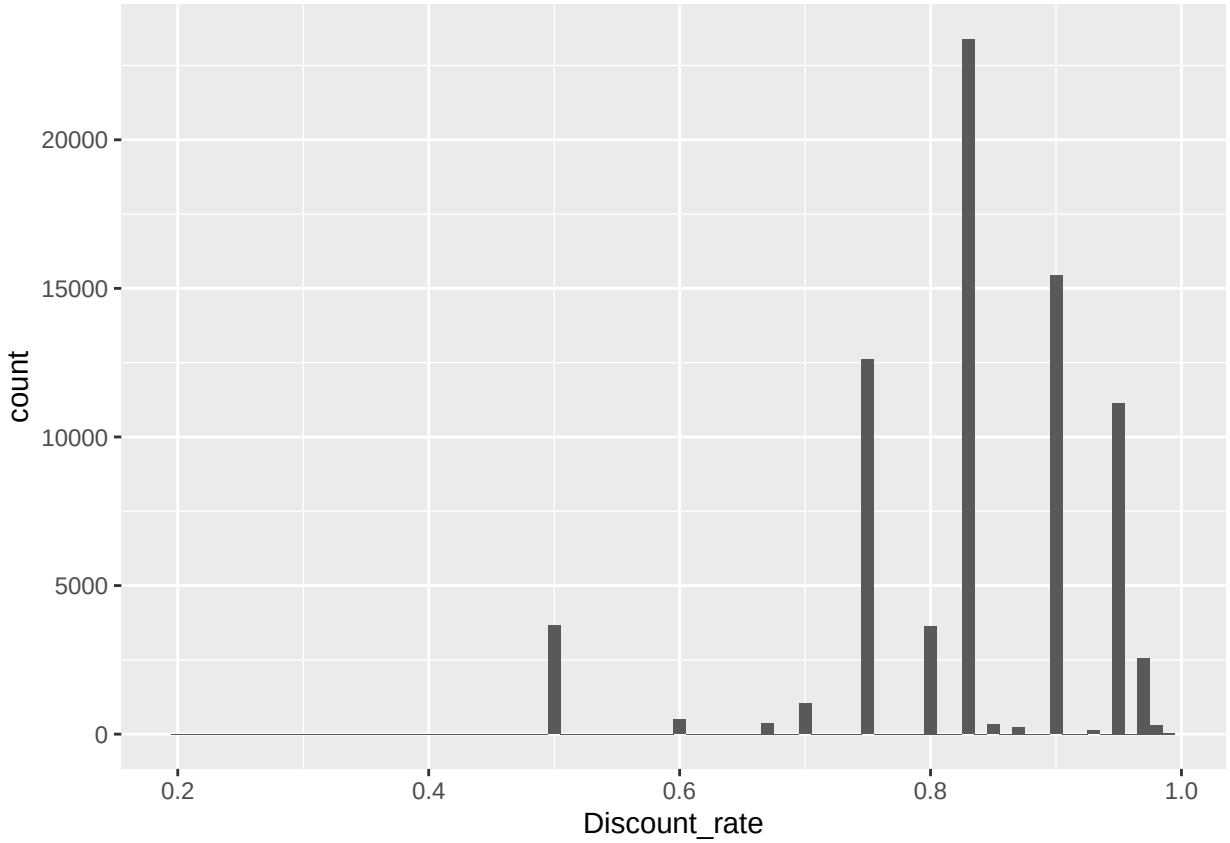
I assigned the variable of "distance_df" by only selecting "Distance", "Date_received" and "Date" not "null. This variable only has the data that the coupon was redeemed by customers. I did a simple plot by looking at the distribution of distance, surprising found out that the coupon redemption is mostly related to distance. As the dataset has illustrated, for"Distance" variable, 0 indicates less than 500 meters;10 indicates more than 5 kilometers. The coupon was redeemed most by people who lives about 500 meters away from the store, as customers live further, they tends less to redeem coupons.

Then I would like to see some correlation between discount rate and coupon redemption.

```
rate_df <- offline_df %>% select("Discount_rate", "Date_received", "Date") %>% filter(Date != "NA")
head(rate_df)
```

```
##   Discount_rate Date_received     Date
## 1          0.95      20160516 20160613
## 2          0.95      20160515 20160521
## 3          0.75      20160321 20160329
## 4          0.83      20160523 20160605
## 5          0.83      20160127 20160221
## 6          0.83      20160207 20160218
```

```
rate_plot <- ggplot(rate_df, aes(x = Discount_rate)) + geom_histogram(binwidth =0.01)
rate_plot
```



According to the graph, as discount rate rises, the coupons tends to have more redemption.

By further investigating the data set, I feel like a model would fit for predicting whether the coupon would likely be redeemed or not based on the discount rate and distance so that business could better initialize what coupon to be distributed to what group of customers.

I'm assuming to change the NA values in Date as "F", while the dates when the coupons was redeemed as "T" to make it a binary and make predictions using discount of rate and distance as a predictor.

```
redeem <- offline_df %>% select("Discount_rate", "Distance", "Date_received", "Date") %>% filter(Date !=
head(redeem)
```

```
##   Discount_rate Distance Date_received     Date
## 1          0.95        0      20160516 20160613
## 2          0.95        0      20160515 20160521
## 3          0.75        0      20160321 20160329
## 4          0.83        0      20160523 20160605
## 5          0.83        0      20160127 20160221
## 6          0.83        0      20160207 20160218
```

```
redeem$Date <- "T"
head(redeem)
```

```
##   Discount_rate Distance Date_received Date
## 1          0.95        0      20160516    T
## 2          0.95        0      20160515    T
```

```
## 3          0.75        0      20160321    T
## 4          0.83        0      20160523    T
## 5          0.83        0      20160127    T
## 6          0.83        0      20160207    T
```

```
not_redeem <- offline_df %>% select("Discount_rate", "Distance", "Date_received", "Date")
not_redeem_replace <- not_redeem[is.na(not_redeem$Date), ]
not_redeem_replace$Date <- "F"
head(not_redeem_replace)
```

```
##   Discount_rate Distance Date_received Date
## 2          0.87        1      20160528    F
## 3          0.95        0      20160217    F
## 4          0.95        0      20160319    F
## 5          0.95        0      20160613    F
## 8          0.90        0      20160429    F
## 9          0.90        1      20160129    F
```

```
new_data <- rbind(redeem, not_redeem_replace)
head(new_data)
```

```
##   Discount_rate Distance Date_received Date
## 1          0.95        0      20160516    T
## 2          0.95        0      20160515    T
## 3          0.75        0      20160321    T
## 4          0.83        0      20160523    T
## 5          0.83        0      20160127    T
## 6          0.83        0      20160207    T
```

I'm going to utilize machine learning skill to train this model. First I'm going to split the new_data into train and test.

```
split <- sample.split(new_data, SplitRatio = 0.8)
split
```

```
## [1] FALSE  TRUE  TRUE  TRUE
```

```
train <- subset(new_data, split == "TRUE")
test <- subset(new_data, split == "FALSE")

redeem_model <- glm(as.factor(Date) ~ Discount_rate+Distance, data = train, family = "binomial")
summary(redeem_model)
```

```
##
## Call:
## glm(formula = as.factor(Date) ~ Discount_rate + Distance, family = "binomial",
##     data = train)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.150652   0.037722  -30.50   <2e-16 ***
```

```
## Discount_rate -1.211503   0.044830  -27.02   <2e-16 ***
## Distance      -0.126017   0.001261  -99.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 407109  on 789960  degrees of freedom
## Residual deviance: 393974  on 789958  degrees of freedom
## AIC: 393980
##
## Number of Fisher Scoring iterations: 6
```

Luckily, I have the model successfully showing above. As the model showing, both distance and discount rate have significantly affected the redemption of a coupon.

Then I would like to test the accuracy of my mode.

```
res <- predict(redeem_model, test, type="response")
head(res)
```

```
##         1         5         9        13        17        21
## 0.09099224 0.10375336 0.14724137 0.10375336 0.03178773 0.08900787
```

```
res <- predict(redeem_model, train, type="response")
head(res)
```

```
##         2         3         4         6         7         8
## 0.09099224 0.11311815 0.10375336 0.10375336 0.10718203 0.11311815
```

```
confmatrix <- table(Actual_value = train$Date, Predicted_value = res > 0.1)
confmatrix
```

```
##             Predicted_value
## Actual_value  FALSE   TRUE
##            F 587496 145929
##            T  32233  24303
```

```
(confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)
```

```
## [1] 0.7744673
```

Ultimately, by adjusting the predicted value threshold to 0.1, I had the accuracy of the model set to work. We could see the accuracy of the model is about 77%! The model is moderately a good fit of the data set.

To summarize, I had demonstrated some data transformation to 3 data sets to investigate how coupons affecting customers and business. I found that the coupon redemption rate is significantly related to discount rate and distance customers live from the stores.

Referances

Epstein L. (Dec 2022) *Advantages and Disadvantages of Using Coupons for Your Business* https://www.investopedia.com/articles/personal-finance/051815/pros-cons-using-coupons-your-business.asp

coupon_pic https://www.liveabout.com/creating-coupon-promotions-2890270

Data

Coupons.csv https://www.kaggle.com/datasets/vysakhvms/coupons

customer_demographics.csv https://www.kaggle.com/datasets/vasudeva009/predicting-coupon-redemption

customer_transaction_data.csv https://www.kaggle.com/datasets/vasudeva009/predicting-coupon-redemption

offline_train.csv https://tianchi.aliyun.com/dataset/137322?t=1690756146750