DM Analysis

Yuka Chen

DM Analysis

Objective

This project is to understand the target audiences to understand what can contribute the reason audiences didn't fill up the information though they have scanned the QR code. Since the limited data size, I combined all data who have scanned the QR code.

Limitation

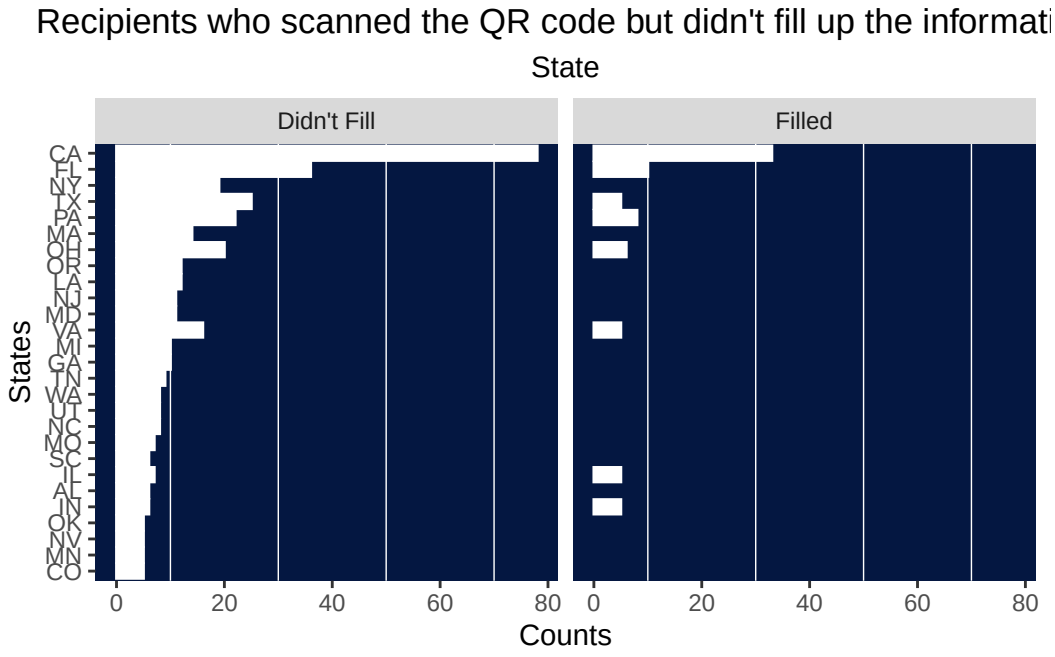
As the sources for June-July used three different, it was quite challenging to clean/combine the data. Moreover, it seems the most `utm_id` do not match with our Snowball_ID.

State	Didn't Fill	Filled	total
CA	78	33	111
FL	36	10	46
TX	25	5	30
PA	22	8	30
OH	20	6	26
NY	19	2	21
VA	16	5	21
MA	14	NA	NA
LA	12	2	14
OR	12	NA	NA
MD	11	1	12
NJ	11	1	12
GA	10	4	14
MI	10	3	13
TN	9	4	13
NC	8	4	12
UT	8	2	10
WA	8	4	12
IL	7	5	12
MO	7	2	9
AL	6	2	8
IN	6	5	11
SC	6	2	8
CO	5	4	9
MN	5	2	7
NV	5	2	7
OK	5	NA	NA
AR	4	1	5
AZ	4	NA	NA
DC	4	2	6
DE	3	NA	NA
KS	3	NA	NA
MT	3	NA	NA
NE	3	NA	NA
WV	2	1	3
CT	1	NA	NA
IA	1	NA	NA
KY	1	NA	NA
MS	1	3	4
ND	1	NA	NA
NM	1	NA	NA
Unknown	1	NA	NA
VT	1	NA	NA
WI	1	2	3
WY	1	NA	NA
ME	NA	1	NA
Total	417	128	489

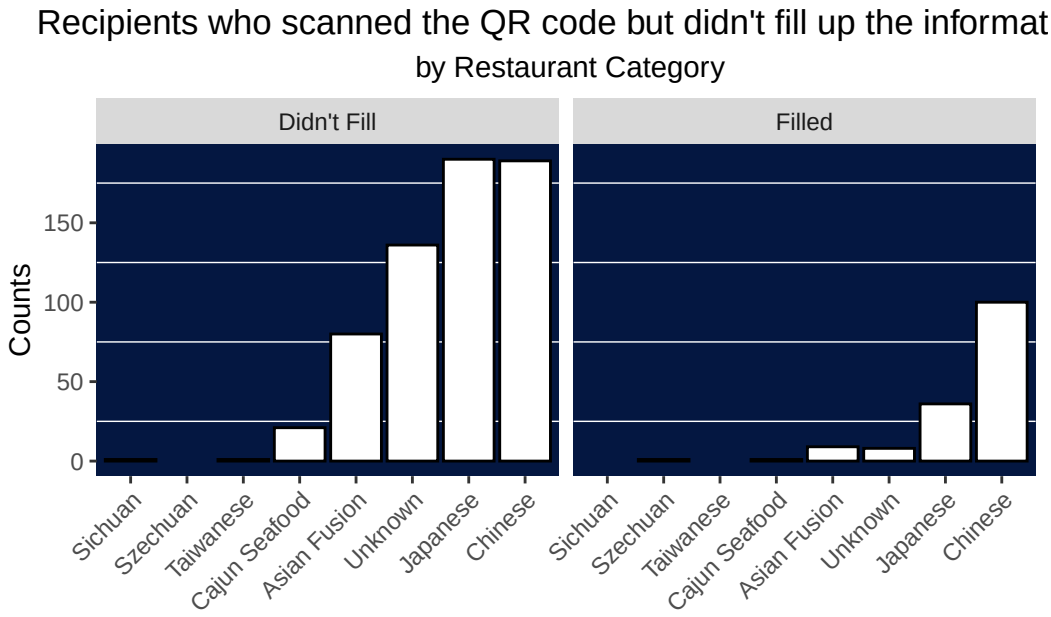
New_Category	Didn't Fill	Filled	total
Chinese	189	100	289
Japanese	190	36	226
Asian Fusion	80	9	89
Unknown	136	8	144
Cajun Seafood	21	1	22
Szechuan	NA	1	NA
Sichuan	1	NA	NA
Taiwanese	1	NA	NA
Total	618	155	770

Data Summary Table and Visualization

State (* - find out the default)

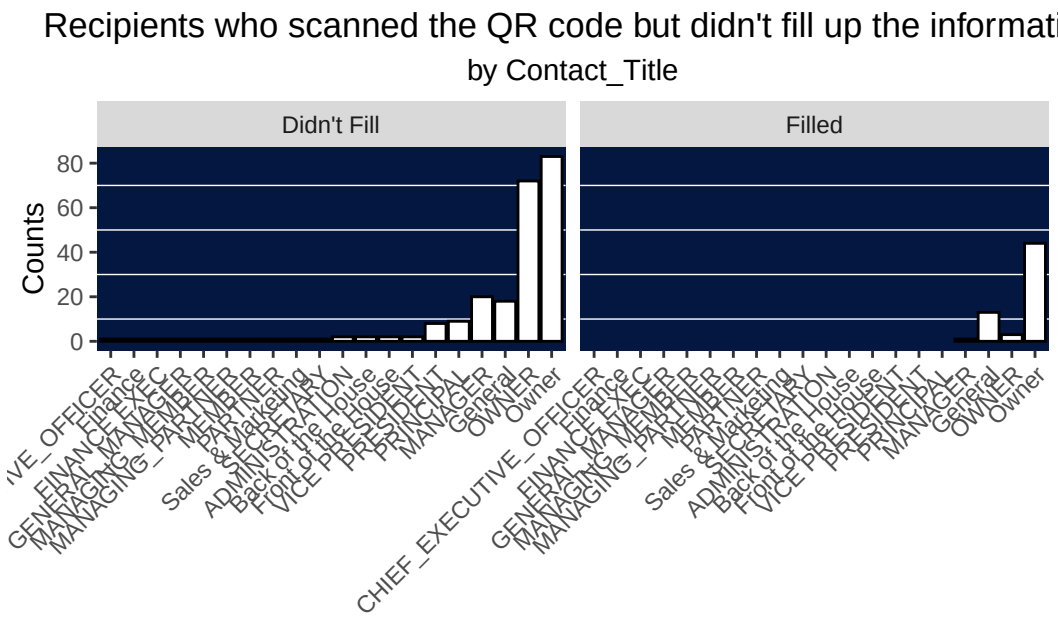


Categorty



By plot, we can see that Cajun Seafood restaurants scanned the QR code, but didn't fill the information. Possibility is that Cajun Seafood restaurants has no ne

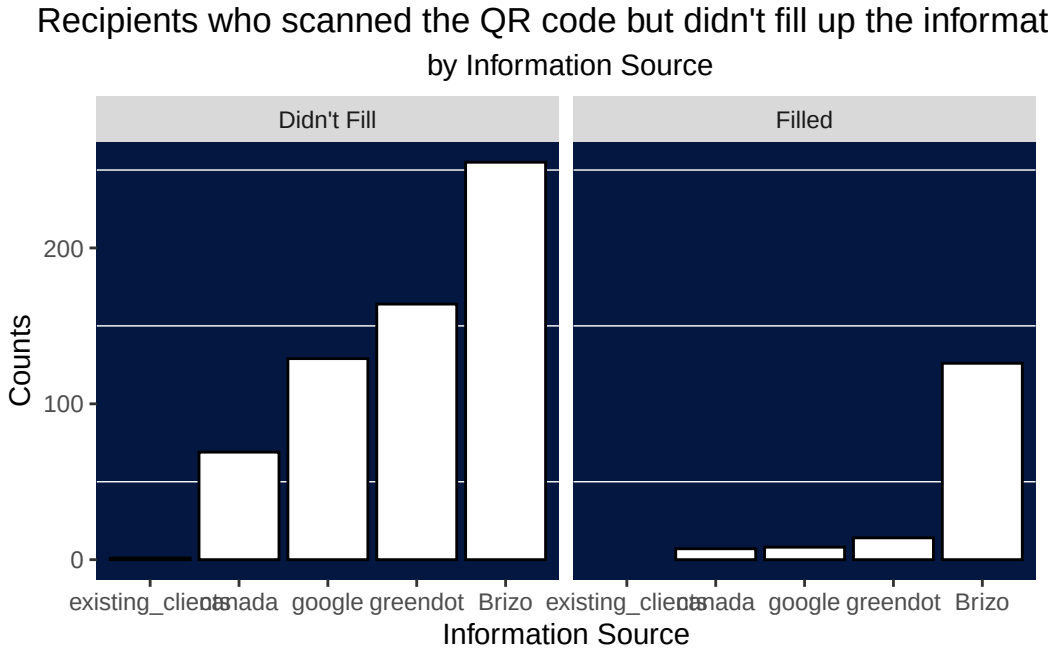
Contact_Title

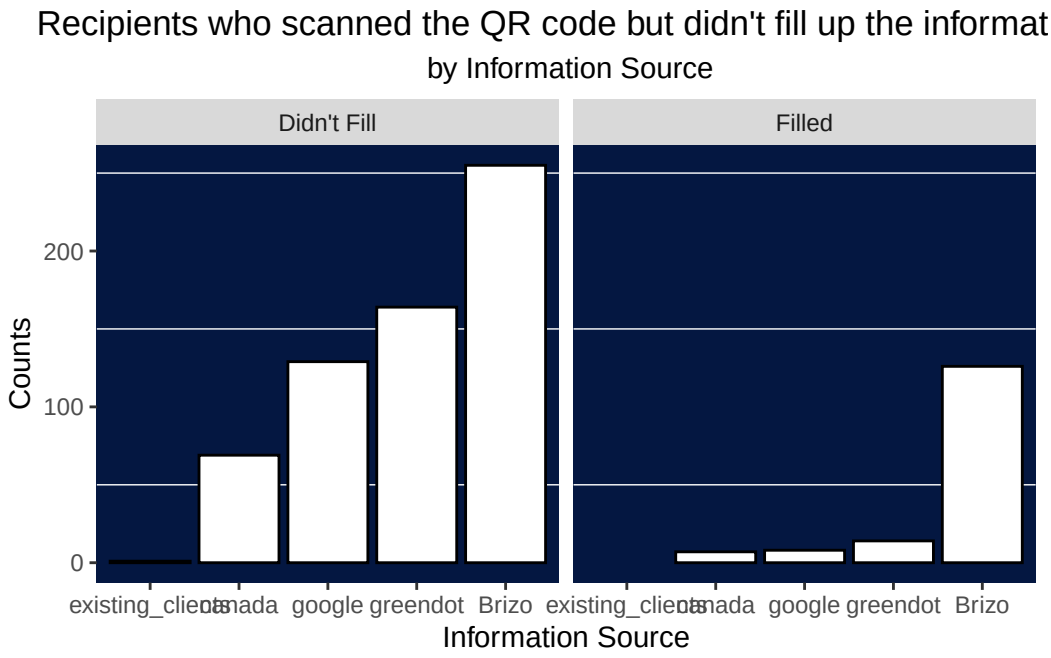


Contact_Title	Didn't Fill	Filled	total
Unknown	390	94	484
Owner	83	44	127
General	18	13	31
OWNER	72	3	75
MANAGER	20	1	21
ADMINISTRATION	2	NA	NA
Back of the House	2	NA	NA
CHIEF_EXECUTIVE_OFFICER	1	NA	NA
FINANCE EXEC	1	NA	NA
Finance	1	NA	NA
Front of the House	2	NA	NA
GENERAL_MANAGER	1	NA	NA
MANAGING_MEMBER	1	NA	NA
MANAGING_PARTNER	1	NA	NA
MEMBER	1	NA	NA
PARTNER	1	NA	NA
PRESIDENT	8	NA	NA
PRINCIPAL	9	NA	NA
SECRETARY	1	NA	NA
Sales & Marketing	1	NA	NA
VICE PRESIDENT	2	NA	NA
Total	618	155	738

InfoSource	Didn't Fill	Filled	total
Brizo	255	126	381
greendot	164	14	178
google	129	8	137
canada	69	7	76
existing_clients	1	NA	NA
Total	618	155	772

Information Source





Models

I ran logistic regression to predict which group are more likely to fill up the information. Unfortunately I had issues when I run a model with all variables and wasn't able to debug it. So instead I ran model separately. The downside of doing it is that we can not consider the possibilities of when all information happened overall.

predicting by State

```
Call:
glm(formula = fill_info_num ~ State, family = binomial(logit),
    data = test)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.07944    0.53033  -3.921 8.82e-05 ***
StateAL         0.98083    0.97361   1.007   0.3137
StateAR         0.69315    1.23744   0.560   0.5754
StateAZ       -15.48663 1978.09024  -0.008   0.9938
StateBC         0.05129    0.71244   0.072   0.9426
StateCA         1.21924    0.56954   2.141   0.0323 *
```

```
StateCO         1.85630    0.85513   2.171   0.0299 *
StateCT       -15.48663 3956.18036  -0.004   0.9969
StateDC         1.38629    1.01550   1.365   0.1722
StateDE       -15.48663 2284.10184  -0.007   0.9946
StateFL         0.79851    0.63955   1.249   0.2118
StateGA         1.16315    0.79451   1.464   0.1432
StateIA       -15.48663 3956.18036  -0.004   0.9969
StateIL         1.74297    0.79000   2.206   0.0274 *
StateIN         1.89712    0.80493   2.357   0.0184 *
StateKS       -15.48663 2284.10184  -0.007   0.9946
StateKY       -15.48663 3956.18036  -0.004   0.9969
StateLA         0.28768    0.92983   0.309   0.7570
StateMA       -15.48663 1057.33380  -0.015   0.9883
StateMB        -0.11778    1.17998  -0.100   0.9205
StateMD        -0.31845    1.17139  -0.272   0.7857
StateME        19.64551 3956.18036   0.005   0.9960
StateMI         0.87547    0.84533   1.036   0.3004
StateMN         1.16315    0.99058   1.174   0.2403
StateMO         0.82668    0.96130   0.860   0.3898
StateMS         3.17805    1.27066   2.501   0.0124 *
StateMT       -15.48663 2284.10184  -0.007   0.9946
StateNB         0.98083    1.27066   0.772   0.4402
StateNC         1.38629    0.81009   1.711   0.0870 .
StateND       -15.48663 3956.18036  -0.004   0.9969
StateNE       -15.48663 2284.10184  -0.007   0.9946
StateNJ        -0.31845    1.17139  -0.272   0.7857
StateNL       -15.48663 1978.09024  -0.008   0.9938
StateNM       -15.48663 3956.18036  -0.004   0.9969
StateNS       -15.48663 1495.29571  -0.010   0.9917
StateNV         1.16315    0.99058   1.174   0.2403
StateNY        -0.17185    0.91317  -0.188   0.8507
StateOH         0.87547    0.70563   1.241   0.2147
StateOK       -15.48663 1769.25771  -0.009   0.9930
StateON         0.10789    0.62034   0.174   0.8619
StateOR       -15.48663 1142.05101  -0.014   0.9892
StatePA         1.06784    0.67209   1.589   0.1121
StateQC        -1.01160    1.15183  -0.878   0.3798
StateSC         0.98083    0.97361   1.007   0.3137
StateSK         1.51983    0.82104   1.851   0.0642 .
StateTN         1.26851    0.80147   1.583   0.1135
StateTX         0.47000    0.72198   0.651   0.5150
StateUnknown  -15.48663 3956.18036  -0.004   0.9969
StateUT         0.69315    0.95197   0.728   0.4665
```

```
StateVA         0.91629    0.73739   1.243   0.2140
StateVT       -15.48663 3956.18036  -0.004   0.9969
StateWA         1.38629    0.81009   1.711   0.0870 .
StateWI         2.77259    1.33463   2.077   0.0378 *
StateWV         1.38629    1.33463   1.039   0.2989
StateWY       -15.48663 3956.18036  -0.004   0.9969
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 774.73  on 772  degrees of freedom
Residual deviance: 691.51  on 718  degrees of freedom
AIC: 801.51

Number of Fisher Scoring iterations: 16
```

From the model to predict a person filling information by their state, we can see that people in California, Colorado, Illinos, Indiana, and WI(Wayomin?) are more likely to fill in information than people living in other states,

predicting by Established Years

```
Call:
glm(formula = fill_info_num ~ Year_Established, family = binomial(logit),
    data = test)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      11.15029   23.27560   0.479    0.632
Year_Established -0.00585    0.01162  -0.504    0.615

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 221.04  on 168  degrees of freedom
Residual deviance: 220.79  on 167  degrees of freedom
  (604 observations deleted due to missingness)
AIC: 224.79

Number of Fisher Scoring iterations: 4
```

The result shows the restaurant established years are not statistically significant for our data (target audiences). It could due to lack of data points.

predicting by Employee_Size (*)

```
Call:
glm(formula = fill_info_num ~ Employee_Size, family = binomial(logit),
    data = test)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.02743    0.43662  -4.643 3.43e-06 ***
Employee_Size -0.04973    0.04939  -1.007    0.314
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 132.16  on 235  degrees of freedom
Residual deviance: 130.59  on 234  degrees of freedom
  (537 observations deleted due to missingness)
AIC: 134.59

Number of Fisher Scoring iterations: 6
```

predicting by restaurant types

```
Call:
glm(formula = fill_info_num ~ New_Category, family = binomial(logit),
    data = test)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.1848     0.3516  -6.214 5.16e-10 ***
New_CategoryCajun Seafood -0.8597     1.0822  -0.794    0.427
New_CategoryChinese        1.5482     0.3727   4.154 3.27e-05 ***
New_CategoryJapanese       0.5213     0.3958   1.317    0.188
New_CategorySichuan      -12.3813   882.7434  -0.014    0.989
New_CategorySzechuan      16.7509   882.7434   0.019    0.985
```

```
New_CategoryTaiwanese     -12.3813   882.7434  -0.014    0.989
New_CategoryUnknown        -0.6484     0.5059  -1.282    0.200
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 774.73  on 772  degrees of freedom
Residual deviance: 699.21  on 765  degrees of freedom
AIC: 715.21

Number of Fisher Scoring iterations: 13
```

The result shows people who owns/manages the Asian Fusion and Chinese restaurants are more likely to fill the information than other types of restaurants.

```
Call:
glm(formula = fill_info_num ~ Contact_Title, family = binomial(logit),
    data = test)

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -1.757e+01  2.797e+03  -0.006    0.995
Contact_TitleBack of the House    6.217e-08  3.956e+03   0.000    1.000
Contact_TitleCHIEF_EXECUTIVE_OFFICER  6.216e-08  4.845e+03   0.000    1.000
Contact_TitleFinance              6.248e-08  4.845e+03   0.000    1.000
Contact_TitleFINANCE EXEC         6.199e-08  4.845e+03   0.000    1.000
Contact_TitleFront of the House   6.176e-08  3.956e+03   0.000    1.000
Contact_TitleGeneral              1.724e+01  2.797e+03   0.006    0.995
Contact_TitleGENERAL_MANAGER      6.243e-08  4.845e+03   0.000    1.000
Contact_TitleMANAGER              1.457e+01  2.797e+03   0.005    0.996
Contact_TitleMANAGING_MEMBER      6.181e-08  4.845e+03   0.000    1.000
Contact_TitleMANAGING_PARTNER     6.199e-08  4.845e+03   0.000    1.000
Contact_TitleMEMBER               6.196e-08  4.845e+03   0.000    1.000
Contact_TitleOwner                1.693e+01  2.797e+03   0.006    0.995
Contact_TitleOWNER                1.439e+01  2.797e+03   0.005    0.996
Contact_TitlePARTNER              6.196e-08  4.845e+03   0.000    1.000
Contact_TitlePRESIDENT            6.196e-08  3.128e+03   0.000    1.000
Contact_TitlePRINCIPAL            6.196e-08  3.093e+03   0.000    1.000
Contact_TitleSales & Marketing    6.197e-08  4.845e+03   0.000    1.000
Contact_TitleSECRETARY            6.192e-08  4.845e+03   0.000    1.000
```

```
Contact_TitleUnknown                  1.614e+01  2.797e+03   0.006    0.995
Contact_TitleVICE PRESIDENT           6.197e-08  3.956e+03   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 774.73  on 772  degrees of freedom
Residual deviance: 715.81  on 752  degrees of freedom
AIC: 757.81

Number of Fisher Scoring iterations: 16
```

It represents that target audience's role do not have any relationship with filling the information.

```
Call:
glm(formula = fill_info_num ~ Campaign, family = binomial(logit),
    data = test)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.4587     0.1935 -12.708   <2e-16 ***
CampaignFree Trail     -0.5617     0.7495  -0.749    0.454
CampaignMooncake        2.1813     0.2277   9.580   <2e-16 ***
CampaignNew Restaurant -14.1074   489.8051  -0.029    0.977
CampaignSummer          -0.8184     0.7459  -1.097    0.273
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 774.73  on 772  degrees of freedom
Residual deviance: 623.31  on 768  degrees of freedom
AIC: 633.31

Number of Fisher Scoring iterations: 15
```

The model shows us that free trail (July) amd mooncake (August) are more successful than other campaign. The reason causes this could be 1) hard to track down the ID (utmID do not match Snowball_ID). 2) simply the campaign meets the audience's need more

Mooncake - offer

```
Call:
glm(formula = fill_info_num ~ group, family = binomial(logit),
    data = mooncake_campaign)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.30010    0.17033  -1.762   0.0781 .
groupOffer_B  0.04521    0.24009   0.188   0.8506
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 386.93  on 282  degrees of freedom
Residual deviance: 386.89  on 281  degrees of freedom
  (31 observations deleted due to missingness)
AIC: 390.89

Number of Fisher Scoring iterations: 4
```

For the Offer A and B, it seems that Offer A is more statistical significant than Offer B. In other words, we can say Offer A has statistical evidence to show it gives some kind of influences on people's deciding to fill the information.

multi variable models (nothing)

```
Call:
glm(formula = fill_info_num ~ State + Year_Established + Employee_Size,
    family = binomial(logit), data = test)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.557e+01  2.313e+07       0        1
StateDC           4.061e-14  3.529e+05       0        1
StateFL           4.146e-15  2.984e+05       0        1
StateGA          -3.568e-14  2.834e+05       0        1
StateIL          -3.248e-14  3.427e+05       0        1
StateMA          -3.675e-14  2.215e+05       0        1
```

```
StateMN          -1.355e-14  3.455e+05       0        1
StateNJ           5.113e+01  2.788e+05       0        1
StateNV          -3.182e-15  3.276e+05       0        1
StateNY          -3.710e-14  2.179e+05       0        1
StateOH           5.113e+01  2.873e+05       0        1
StateTX           5.413e-15  3.570e+05       0        1
StateUT          -1.945e-15  3.102e+05       0        1
Year_Established -1.210e-15  1.160e+04       0        1
Employee_Size    -2.886e-16  1.964e+04       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.3768e+01  on 23  degrees of freedom
Residual deviance: 3.7848e-10  on  9  degrees of freedom
  (749 observations deleted due to missingness)
AIC: 30

Number of Fisher Scoring iterations: 24
```