DM Analysis

Yuka Chen

Marketing Operation Inter @ Wonders

Direct Mail Analysis

1. Objective

The objective of this project is to understand the target audience's reasons for not filling out the information after scanning the QR code. Since the data set is limited in size for each campaign, I have combined all data from all campaign who have scanned the QR code and group based on they fill the information form or not.

2. Limitation

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

3. Highlights of the report

- People in California, Colorado, Illinois, Indiana, and Wisconsin are more likely to fill in information than people living in other states.
- Individuals who own or manage Asian Fusion and Chinese restaurants are more likely to fill in the information than those associated with other types of restaurants.
- The role of the target audience does not have any significant relationship with filling in the information.

4. Take

People in certain demography do have some kind of relationship to do with their decision on filling the information after they scanned our QR code.

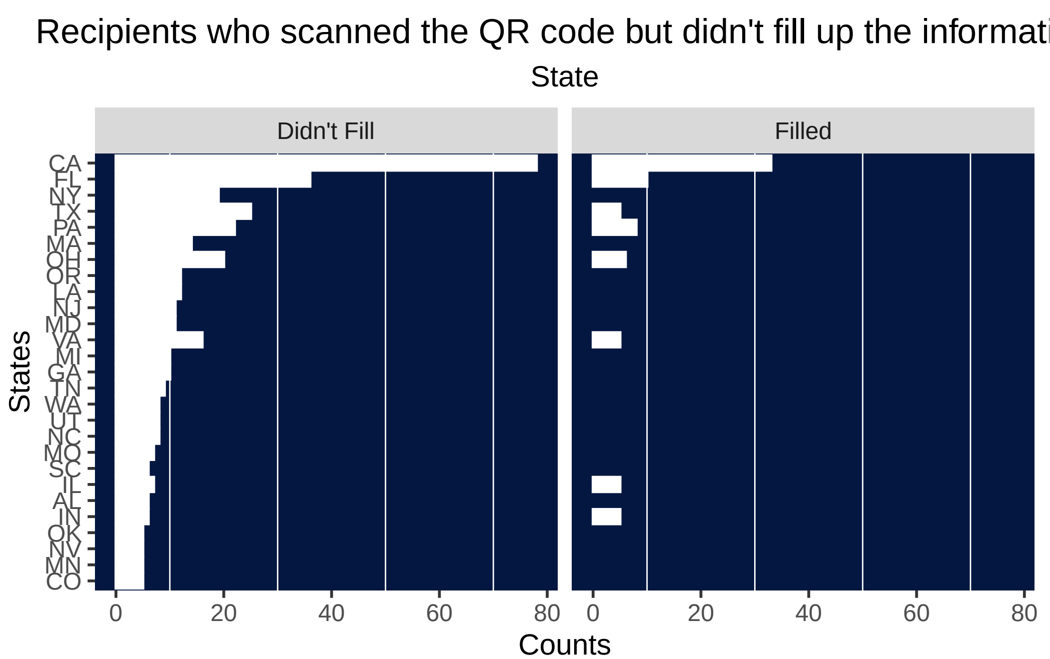
5. What's Next?

- **Find out the data transformation discrepancy (why not match) so we can predict clients behavior better**
- Run other machine learning model to see if we can get better predictions
- Run the prediction models by campaigns (i.e. June/July audience info contains Gender, Owner Ethnicity)
- Clean/Maintain the data well so we can run a model better - better prediction. We could calculate the possibility of a person to fill the information by their personal background and restaurants information.
- find out pattern on people who didn't scan vs scanned (break the first barrier)

Analysis

6. Data Summary Table and Visualization

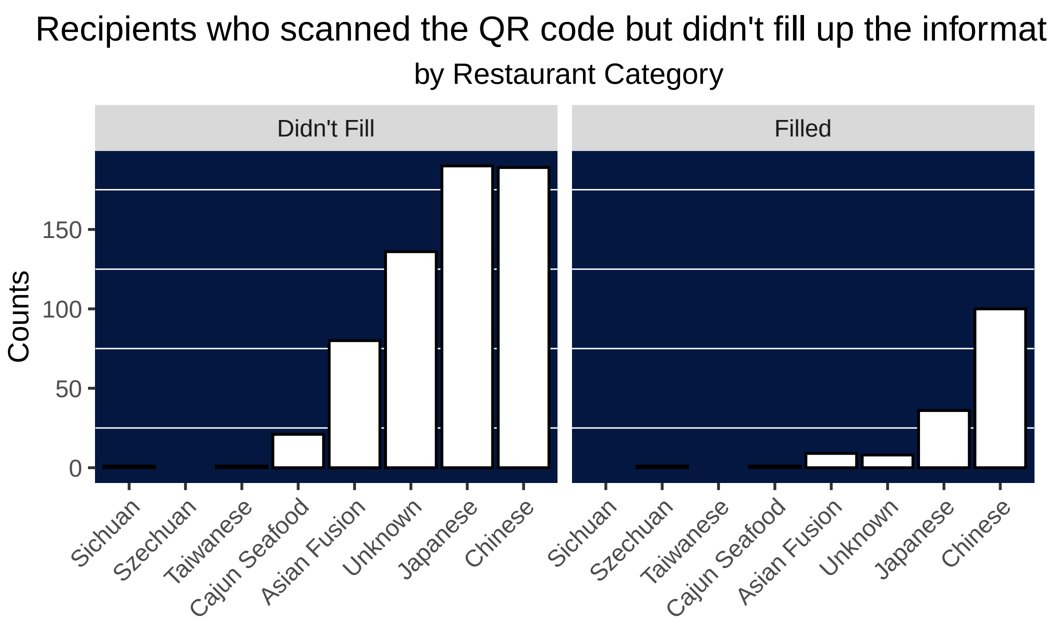
State



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

Categorty

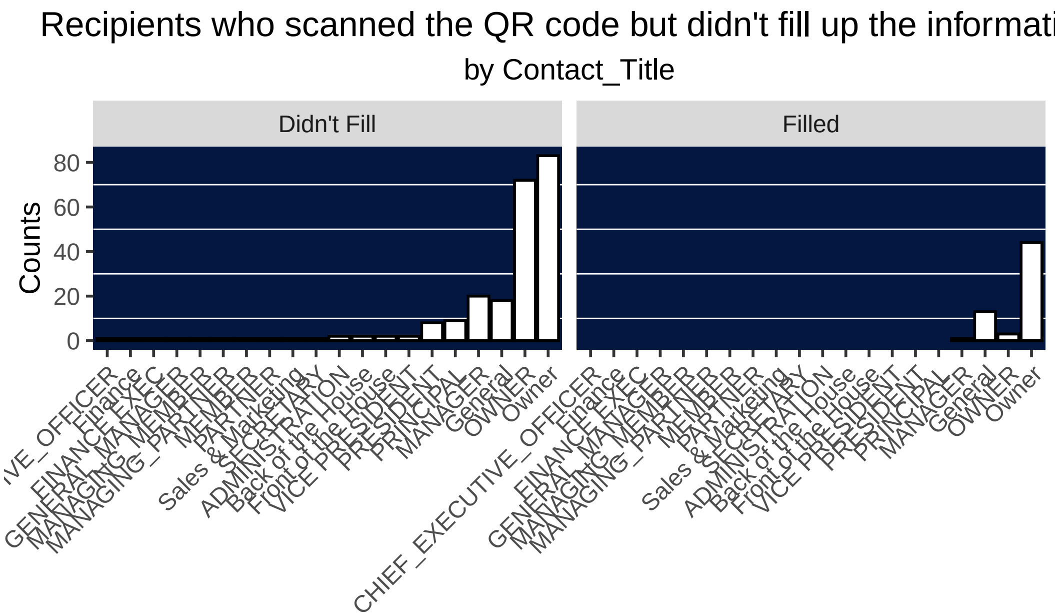


By plot, we can see that Cajun Seafood and Japanese restaurants scanned the QR code, but didn't fill in the information.

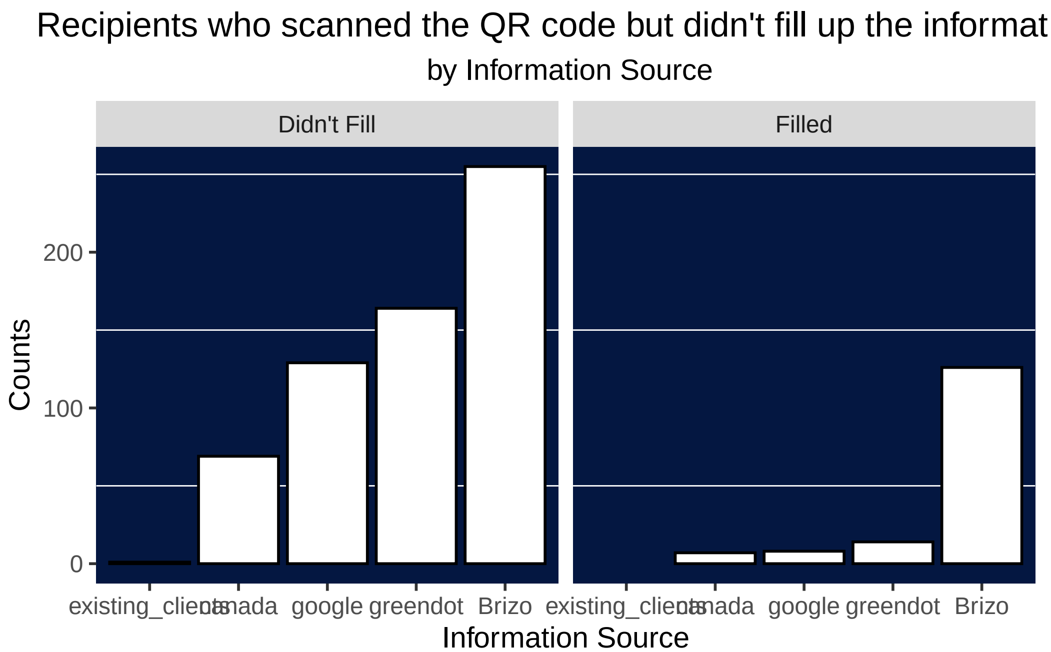
Contact_Title	Didn't Fill	Filled	total
Unknown	390	94	484
Owner	83	44	127
General	18	13	31
OWNER	72	3	75
MANAGER	20	1	21
ADMINISTRATION	2	NA	NA
Back of the House	2	NA	NA
CHIEF_EXECUTIVE_OFFICER	1	NA	NA
FINANCE EXEC	1	NA	NA
Finance	1	NA	NA
Front of the House	2	NA	NA
GENERAL_MANAGER	1	NA	NA
MANAGING_MEMBER	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

Contact_Title



Information Source



7. 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.

7.1 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
```

Interpretation

- **CA, CO, IL, IN, WI has a positive coefficient with a small p-value (<0.05, *), which suggests that people in these states have higher log-odds of the response variable compared to the reference state.**
- The coefficients for each state (e.g., AL, AR, AZ, etc.) indicate how much the likelihood of people filling in information in that specific state differs from the baseline state.
- For example, in the case of CA, it has a positive coefficient of approximately 1.21924, and a p-value of 0.0323 (*), suggesting that people in California are more likely to fill in information after scanning the QR code compared to the baseline state.
- Conversely, for IL, it has a positive coefficient of approximately 1.74297, and a p-value of 0.0274 (*), indicating that people in Illinois are also more likely to fill in information compared to the baseline state.
- MS has a positive coefficient of approximately 3.17805, and a p-value of 0.0124 (*), implying that people in Mississippi have a significantly higher likelihood of filling in information compared to the baseline state.
- CO, IN, NC, and WI also have positive coefficients with p-values less than 0.05 (*), suggesting that people in these states are more likely to fill in information compared to the baseline.

- On the other hand, some states like AZ, CT, DE, IA, KS, KY, NM, ND, ON, OR, TX, UT, VT, and WY have coefficients with p-values greater than 0.05, indicating that there is no statistically significant difference in the likelihood of people filling in information in these states compared to the baseline state.
- Overall, this analysis examines how the state variable relates to the likelihood of people filling in information after scanning a QR code, with positive coefficients suggesting a higher likelihood, and negative coefficients suggesting a lower likelihood, compared to the baseline state (which is not explicitly mentioned in the output).

7.2 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
```

Interpretation

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

7.3 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
```

Interpretation

- The coefficient for Employee_Size is approximately -0.01194, and it has a relatively high p-value (0.74), indicating that there is no statistically significant relationship between the number of employees in a restaurant and the likelihood of people filling in information after scanning the QR code. The p-value is greater than the conventional significance level (0.05), suggesting that the variable "Employee_Size" does not have a significant impact on the outcome.
- Overall, based on this analysis, the number of employees in a restaurant does not appear to be a significant predictor of whether people will fill in information after scanning the QR code, at least when considering data excluding cases.

7.4 predicting by restaurant types

```
Call:
glm(formula = fill_info_num ~ New_Category, family = binomial(logit),
    data = test[test$New_Category != "Unknown", ])

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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 683.99  on 628  degrees of freedom
Residual deviance: 637.42  on 622  degrees of freedom
AIC: 651.42

Number of Fisher Scoring iterations: 13
```

Interpretation

- In this logistic regression analysis, the "Intercept" represents the baseline likelihood of people filling in information after scanning a QR code when the New_Category is Asian Fusion.
- The coefficient for Asian Fusion is -2.1848, and it has a low p-value (***), indicating that when the restaurant's category is Asian Fusion, there is a statistically significant decrease in the likelihood of people filling in information compared to other types of restaurants.
- The coefficient for Chinese is 1.5482, and it has a low p-value (***), indicating that when the restaurant's category is Chinese, there is a statistically significant increase in the likelihood of people filling in information compared to the baseline "Asian Fusion" category.
- Other New_Category coefficients, such as "Cajun Seafood,Japanese, Sichuan, Szechuan, Taiwanese, and Unknown, do not show a statistically significant influence on the likelihood of people filling in information, as their p-values are higher than the conventional significance level (0.05).
- Overall, the type of restaurant category, specifically "Chinese," appears to be associated with a statistically significant increase in the likelihood of people filling in information after scanning the QR code, compared to the baseline category, "Asian Fusion."

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

```
    data = test[test$Contact_Title != "Unknown", ])

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -1.857e+01  4.612e+03  -0.004    0.997
Contact_TitleBack of the House       -2.157e-07  6.523e+03   0.000    1.000
Contact_TitleCHIEF_EXECUTIVE_OFFICER -2.160e-07  7.989e+03   0.000    1.000
Contact_TitleFinance                 -2.157e-07  7.989e+03   0.000    1.000
Contact_TitleFINANCE EXEC            -2.158e-07  7.989e+03   0.000    1.000
Contact_TitleFront of the House      -2.156e-07  6.523e+03   0.000    1.000
Contact_TitleGeneral                  1.824e+01  4.612e+03   0.004    0.997
Contact_TitleGENERAL_MANAGER         -2.156e-07  7.989e+03   0.000    1.000
Contact_TitleMANAGER                  1.557e+01  4.612e+03   0.003    0.997
Contact_TitleMANAGING_MEMBER         -2.157e-07  7.989e+03   0.000    1.000
Contact_TitleMANAGING_PARTNER        -2.157e-07  7.989e+03   0.000    1.000
Contact_TitleMEMBER                  -2.157e-07  7.989e+03   0.000    1.000
Contact_TitleOwner                    1.793e+01  4.612e+03   0.004    0.997
Contact_TitleOWNER                    1.539e+01  4.612e+03   0.003    0.997
Contact_TitlePARTNER                 -2.157e-07  7.989e+03   0.000    1.000
Contact_TitlePRESIDENT               -2.157e-07  5.157e+03   0.000    1.000
Contact_TitlePRINCIPAL               -2.157e-07  5.099e+03   0.000    1.000
Contact_TitleSales & Marketing       -2.157e-07  7.989e+03   0.000    1.000
Contact_TitleSECRETARY               -2.157e-07  7.989e+03   0.000    1.000
Contact_TitleVICE PRESIDENT          -2.157e-07  6.523e+03   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 297.89  on 288  degrees of freedom
Residual deviance: 239.28  on 269  degrees of freedom
AIC: 279.28

Number of Fisher Scoring iterations: 17
```

- The *Intercept* represents the baseline likelihood of filling in information when all other factors are zero. It's not statistically significant, indicating that the baseline likelihood is not significantly different from zero.
- The coefficients for different Contact_Title categories show how each category influences the likelihood of filling in information compared to a reference category (which is not explicitly mentioned). However, the p-values for most coefficients are very high, indicating that these categories don't significantly impact the likelihood of filling in information.
- **Overall, the Contact_Title variable, as included in the model, doesn't appear to have a statistically significant influence on whether people fill in**

information after scanning the QR code.

```
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
```

Interpretation

- The model used Iphone campaign as a base line, meaning the intercept, and the estimated coefficient is -2.4587. It is statistically significant (p-value < 0.001), suggesting that this campaign has a significant negative impact on the log-odds of fill_info_num. As the coefficient is negative, being part of the Iphone campaign is associated with lower odds of fill_info_num being 1.
- For the Free Trail campaign, the estimated coefficient is -0.5617. However, it is not statistically significant (p-value = 0.454), indicating that this campaign does not have a significant impact on the log-odds of fill_info_num.
- **For the Mooncake campaign, the estimated coefficient is 2.1813. It is statistically significant (p-value < 0.001), suggesting that this campaign has a significant positive impact on the log-odds of fill_info_num. As the coefficient is positive, being part of the Mooncake campaign is associated with higher odds of fill_info_num being 1.**

- For the New Restaurant campaign, the estimated coefficient is -14.1074. However, it is not statistically significant (p-value = 0.977), indicating that this campaign does not have a significant impact on the log-odds of fill_info_num.
- For the Summer campaign, the estimated coefficient is -0.8184. However, it is not statistically significant (p-value = 0.273), indicating that this campaign does not have a significant impact on the log-odds of fill_info_num.

Overall, the Mooncake campaign appears to have a significant positive impact on the likelihood of people filling the information form (= fill_info_num being 1), while the other campaigns do not seem to have a significant effect based on their p-values